Characterisation of plosive, fricative and aspiration components in speech production

by

Philip J.B. Jackson

Thesis submitted for the degree of Doctor of Philosophy

to the

Faculty of Engineering and Applied Science

Department of Electronics and Computer Science

May 2000

Supervised by Dr. Christine H. Shadle

Examined by Dr. R.I. Damper and Prof. D.M. Howard

University of Southampton

ABSTRACT

Doctor of Philosophy

Characterisation of plosive, fricative and aspiration components in speech production

by Philip J.B. Jackson

This thesis is a study of the production of human speech sounds by acoustic modelling and signal analysis. It concentrates on sounds that are not produced by voicing (although that may be present), namely plosives, fricatives and aspiration, which all contain noise generated by flow turbulence. It combines the application of advanced speech analysis techniques with acoustic flow-duct modelling of the vocal tract, and draws on dynamic magnetic resonance image (dMRI) data of the pharyngeal and oral cavities, to relate the sounds to physical shapes.

Having superimposed vocal-tract outlines on three sagittal dMRI slices of an adult male subject, a simple description of the vocal tract suitable for acoustic modelling was derived through a sequence of transformations. The vocal-tract acoustics program VOAC, which relaxes many of the assumptions of conventional plane-wave models, incorporates the effects of net flow into a one-dimensional model (viz., flow separation, increase of entropy, and changes to resonances), as well as wall vibration and cylindrical wavefronts. It was used for synthesis by computing transfer functions from sound sources specified within the tract to the far field.

Being generated by a variety of aero-acoustic mechanisms, unvoiced sounds are somewhat varied in nature. Through analysis that was informed by acoustic modelling, resonance and anti-resonance frequencies of ensemble-averaged plosive spectra were examined for the same subject, and their trajectories observed during release. The anti-resonance frequencies were used to compute the place of occlusion.

In vowels and voiced fricatives, voicing obscures the aspiration and frication components. So, a method was devised to separate the voiced and unvoiced parts of a speech signal, the pitch-scaled harmonic filter (PSHF), which was tested extensively on synthetic signals. Based on a harmonic model of voicing, it outputs harmonic and anharmonic signals appropriate for subsequent analysis as time series or as power spectra. By applying the PSHF to sustained voiced fricatives, we found that, not only does voicing modulate the production of frication noise, but that the timing of pulsation cannot be explained by acoustic propagation alone.

In addition to classical investigation of voiceless speech sounds, VOAC and the PSHF demonstrated their practical value in helping further to characterise plosion, frication and aspiration noise. For the future, we discuss developing VOAC within an articulatory synthesiser, investigating the observed flow-acoustic mechanism in a dynamic physical model of voiced frication, and applying the PSHF more widely in the field of speech research.

Acknowledgements

I would like to express my gratitude to Dr. Christine Shadle for her enduring supervision over the three and a half years, during which time she has adopted the various roles of critic, advisor, subject, sounding board, mentor and colleague. There is little doubt that without her help I would not have been able to gain sufficient access to the field of speech production to make this project meaningful and worthwhile. Aside from the significant task of overseeing the writing of this report, she has supplied experimental data from her own thesis. I have also benefitted from her collaborative work in terms of computer software, medical images and essential sound recording and editing facilities.

For their guidance and advice, my thanks go to Dr. Bob Damper, Prof. Peter Davies, Dr. Paul White, Dr. Mohammad Mohammad and Prof. Phil Nelson, and to Dr. Anna Barney who has kindly worked through much of Appendix A with me. I would also like to acknowledge my examiners, Dr. Bob Damper and Prof. David Howard, and the peer reviewers for their constructive comments on papers deriving from this work. For their patience as subjects, I thank Sharon Benton and Luis-Miguel Teixeira de Jesus. Professor Davies wrote the earlier versions of VOAC which were a key part of the acoustic modelling herein, as were the magnetic resonance images resulting from Mohammad's thesis that acted as source material.

Personally, I am indebted to my family and friends for their support and understanding, and to Maite Villoria-Nolla for the love that she has shown me. Finally, I would like to thank the Faculty of Engineering and Applied Science, and the Department of Electronics and Computer Science for substantial assistance, both financial and practical, not to mention all those with whom I have exchanged ideas during the course of this project, called Nephthys.

This thesis was submitted to the Department of Electronics and Computer Science, University of Southampton in fulfillment of the requirements for the degree of Doctor of Philosophy. It is entirely my own work and, except where otherwise stated, describes my own research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

This study considers the nature of unvoiced sounds in human speech, both their means of production and their acoustic characteristics. Voicing and its acoustic consequences have been studied in great detail, and much is known about the oscillation of the vocal folds and the ensuing propagation of acoustic disturbances. Indeed, the location of the sound source is widely accepted to be at the glottis, and the acoustic principles of how the sound is modified by the vocal tract are also well understood. Yet, speech contains many other sounds whose source location is not so clearly defined and whose interaction with the vocal tract is not well understood. These sounds are the subject of the present study. In particular, we investigate sounds generated by flow turbulence, such as frication and aspiration, and by the sudden release of air, i.e., stop consonants or plosives.

The motivation for this research comes from the difficulty of synthesising natural-sounding speech. One approach would perform analyses of speech signals in order to describe the acoustically-significant features, and another would model the acoustics of the vocal tract in order to assess the influence of source location. We have done both and compared their results. Our specific objective was to derive the theoretical and empirical basis for an improved generalised model of the production of unvoiced sounds, but along the way we have not only developed a widely-applicable speech decomposition tool, but uncovered some interesting information concerning other types of sounds.

There are challenges to be faced with either approach. Unlike voiced speech signals, unvoiced sounds tend not to be deterministic, being generated from air turbulence, which is a kind of stochastic process. The random nature of the signals raises questions not only of how best to analyse them, but what the features are that the analysis should try to uncover. For example, should we be interested in vocal-tract anti-resonances, as well as the formant resonances estimated in traditional analyses? Furthermore, the features of a plosive are not only

predominantly noisy, but highly time variant. The sound termed *aspiration* almost always occurs in the company of other sound sources, particularly voicing, which tends to overshadow the part of the signal in which we are interested. To study the unvoiced component during phonation, we developed a speech processing technique to split the signal into two components: harmonic (representing the voiced part) and anharmonic (representing the unvoiced part).

The exact location of the unvoiced sound sources is also difficult to ascertain. Although the position of the constriction may be well defined in a fricative, the acoustic source derives from the turbulence downstream of the constriction and its properties are enormously dependent on the local geometry. As we shall see, any uncertainty surrounding where breath noise is generated causes difficulty for accurate acoustic modelling. For plosives, the problem is similar, but with the additional complication of transient factors and the associated measurement difficulties. If we want to predict the acoustics of the human airways, adequate measurements are needed along the vocal tract, for which we have resorted to using magnetic resonance imaging (MRI).

Moreover, the presence of voicing influences the production of turbulence and modifies the generation of noise in curious ways. While we have not sought to grapple with the extensive body of aero-acoustic theory, we have gathered empirical results through our innovative analysis method that give strong evidence of aero-acoustic interaction, and suggest a possible mechanism for voiced fricatives.

1.1.1 Purpose

The objective of this thesis is therefore to study speech sounds that involve turbulence, and to try to improve models of their production. The three classes of these sounds are called frication, aspiration and plosion, and they occur in fricatives, breathy voicing, stops and affricates. Current unanswered questions include: where is aspiration noise produced? Is the noise source really weaker in voiced fricatives that in voiced ones? What are the characteristics of any modulation of the noise source, and how can voicing cause this effect? How do the frication and aspiration stages in a stop consonant relate to the noise in a fricative or breathy vowel?

Many of the shortcomings in our present knowledge are the result of difficulties with theoretical models or conversely with analysing speech signals. Progress can be made by applying existing analysis tools to the unvoiced sounds in a new way. For example, we apply ensemble averaging of short-term power spectra to plosive releases in order to capture the characteristic properties of the ensuing sequence of sounds. An alternative route to finding new information about noise signals is to develop a technique that will enable us to explore sounds as we could not before. Mixed-source speech, where both voicing and an unvoiced source are operating, has previously been hard to analyse because of the interference of the two (or more) sources, yet it is of crucial importance to the study of breath noise and essential for many voiced consonants,

including fricatives. In this respect, we have created the pitch-scaled harmonic filter (PSHF) to extract the voiced and unvoiced contributions, which not only opens the way for studying frication and aspiration in the presence of voicing, but offers the opportunity to examine how the different sources interact and perhaps, through doing so, to learn more about the mechanisms by which the turbulence noise is produced.

As already mentioned, the theoretical models may, by their form or their assumptions, fail to describe the rich properties of consonants accurately. In fact, many popular models, such as those embodied in synthesis systems, being originally developed from the results of vowel experiments, lack the sophistication required to encapsulate the behaviour of an aero-acoustic source mechanism. They tend to represent the vocal-tract transfer function (VTTF) as an all-pole filter that only permits plane-wave, acoustic propagation, whereas flow convection is obviously an intrinsic element of turbulence noise generation. As a first step, we have incorporated flow into the acoustic model we use, adapted the model to include non-glottal sources, and investigated the influence of non-acoustic fluid motion on sound production.

1.1.2 Problem statement

The source-filter model is a highly successful description of the speech production process, which has been used for coding, modification and synthesis of speech because of its parsimony. The source and filter elements are attractive since they can be taken to correspond roughly to physical entities. Following Lighthill's acoustic analogy (Lighthill 1952; 1954), we can express the source of acoustic pressure waves as being equivalent to flow monopole, dipole or quadrupole point sources in a uniform medium. The consequences of such compact sources to the sound field at the lips, and thence to the far field depend on the acoustic transfer function from the source to the lips, the plane from which free radiation takes place. This VTTF cannot be computed exactly, but suitable assumptions can lead to estimates of very high accuracy over the frequency range of interest. The effect of the VTTF can be thought of as effectively filtering the source function, and for this reason the paradigm is referred to as the source-filter model. Satisfactory results have been achieved with plane-wave acoustic models, such as Fant's (1960), alternatively represented as a transmission line in the classic electrical analogue (Flanagan 1972). Our model builds on these successes but, as well as relaxing various assumptions, includes a factor of primary importance for consonants — the effect of net flow through the vocal tract (Davies et al. 1993).

Calculating the vocal-tract filter characteristic is an indirect means of estimating a single sound source, which might be achieved via the combination of acoustic modelling and signal analysis. In addition, there are signal processing techniques that have been developed for decomposing speech signals into quasi-periodic and aperiodic components, which can be con-

sidered to be estimates of the voiced and unvoiced parts, respectively. Ideally, the aperiodic or anharmonic component would contain all, and only, the filtered noise sources, and the periodic or harmonic component precisely the vocal-tract-filtered voicing source.

Plosive, fricative and aspiration noise are of critical interest in the realistic production of speech. A better understanding of these can be a help in the diagnosis of pathological speech (e.g., hoarseness, dysphonia) and, of course, improve naturalness in speech synthesis. In particular, there are many questions surrounding aspiration. What is it? When does it occur? Where is it generated? How is the turbulence noise produced? How can we measure it? How can we model it? Our analysis methodology consists of: defining aspiration, making speech recordings, building speech analysis tools, analysing single-source speech (viz. modal vowels, and unvoiced plosives and fricatives), and decomposing mixed-source speech (breathy speech and voiced fricatives).

1.1.3 Applications

Any increment to knowledge of how plosives, fricatives and aspiration noise are produced is likely to have a positive impact on a whole range of activities, not only in fundamental aspects of speech research, but also in applications. There are four principal application areas: science, technology, medicine and education. Speech science includes analysis, production and perception of speech. The invention or development of any speech analysis technique, such as in Chapter 5 of this thesis, may lead to new findings in related fields, as we will see in Chapter 7. For example, a better understanding of the characteristics of speech sounds can inform studies of their perception and, along with appropriate analysis techniques, can be used to assess voice quality: breathy, creaky, harsh, raspy, rough, etc.

Technological applications include articulatory synthesis, concatenative synthesis and the field of speech modification. In synthesis, more natural-sounding, aspirated and breathy speech could be produced using acoustic models developed to account for flow phenomena, as could other voice qualities. Improved descriptions of the plosive and fricative sounds would lead directly to modifications of synthesis models, an example of which occurs in Section 7.5, with resulting benefits. Such descriptions might suggest ways of improving their coding efficiency. Moreover, being based on physical considerations, they may lend themselves more naturally to enhancement and modification tasks, possibly with implications for automatic speech recognition.

Since verbal communication is an intrinsic aspect of human existence, deviations from normal speech performance are of interest to medical specialists. Conversely, medical devices can be a valuable resource for speech research, such as the dMRI scans used in Chapter 3. Tools have been developed to facilitate clinical assessment of patients and the diagnosis of

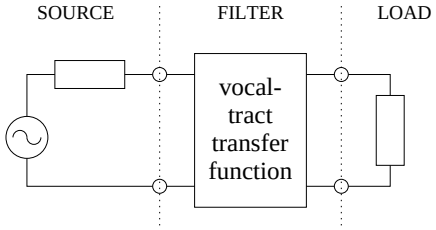


Figure 1.1: Simple depiction of the source-filter model.

pathologies. With knowledge of the mechanisms of production and the noise-flow relation, these systems can be improved. Many techniques developed for speech analysis can be of value in clinical measurements: of harmonics-to-noise ratio (HNR, Yumoto et al. 1982; Muta et al. 1988; Qi and Hillman 1997), which can be estimated dynamically by the PSHF (see Chapter 5); of vocal effort (Richard and d'Alessandro 1997); of phonatory quality (Blomgren et al. 1998), which we examine in Chapter 6.

Specialist speech analysis packages are already used by professional singers to assist in training their voice, yet the potential for speech applications in the education sector are enormous, for instance, books that talk to children and recognize what they say, aids for the handicapped and foreign language teaching software.

1.2 Speech production

The source-filter paradigm involves modelling the human speech production system as a result of a sound source being passed through a filter has proved to be an effective means of explaining many observations, as well as providing a powerful analogy (Fant 1960; Mermelstein 1971; Badin 1991; Shadle 1995a). In its simplest terms, the source-filter model implies that the source and filter elements are independent and it is normally assumed that the filter behaves linearly. The outputs of the filter determine the sound that is radiated into the far field, which acts as a load, as shown in Figure 1.1.

Phonation, or voicing, is the main source of sound generation for the majority of speech, and can be modelled as chaotic behaviour of a non-linear oscillator. It is particularly important for modal vowels, nasals and liquids, where it occurs almost exclusively, but also for other voiced consonants, such as /b, d, g, v, ð, z, ʒ, dʒ/. Its features include amplitude, pitch, jitter, shimmer, roughness and hoarseness, some of which will be discussed later in Chapter 4. The rest of this section describes various other sorts of sound, namely, the products of unvoiced acoustic sources.

1.2.1 Fricatives

Fricatives, such as /f/, /s/ and /ʃ/, as in *fun*, *sun* and *shun*, involve a type of sound source — turbulence noise — that is produced by the jet of air flowing through a constriction in the vocal tract. In general terms, turbulence noise is an aero-acoustic phenomenon that is generated by the fluctuating pressures in turbulent flow conditions. Turbulent flows occur downstream of points of flow separation (typically for Reynolds numbers $Re > 2500$, Davies et al. 1993), for instance in the wake of the jet, such as that formed at the tongue tip during /s/. The impingement of such flows on surfaces, particularly edges, converts the order of the acoustic source from quadrupole to dipole, greatly increasing the efficiency of sound radiation (Lighthill 1954; Curle 1955; Pierce 1981). Within the vocal tract, the teeth can act as such an obstacle, or the source may be distributed with respect to the direction of flow, for instance, along the palate. In speech, this kind of noise generation is called frication. A *frication* source can be considered to be a turbulence-noise source, caused by flow through a supraglottal constriction, and is sometimes enhanced by an edge or obstacle in the path of the jet.

Turbulence noise occurs in a large class of speech sounds: (i) in stationary form, (ii) accompanied by and perhaps modulated by periodic sound, and (iii) in transient form. When the source of the turbulence is localised, near an articulated constriction, the resulting noise is usually called frication, which has been much studied (Stevens 1971; Stevens et al. 1992; Shadle 1985, 1990, 1995a, 1995b; Shadle et al. 1991; Shadle et al. 1992; Badin 1991; Narayanan et al. 1995; Mair and Shadle 1996; Sinder et al. 1998). Plosives, as we shall see in Section 1.2.2, contain a sequence of transient sounds that can be crudely categorised as burst, frication and aspiration, leading up to voice onset. Noise that is not impulsive (like burst noise) or attributable to a supraglottal constriction (like frication) tends to be called *aspiration*, giving it a confusing variety of definitions (see Section 1.2.3).

The production of voiced fricatives comprises two predominant sources of sound exciting the vocal-tract resonances: the phonation source (voicing), produced by vocal-fold oscillation, and the frication noise source, produced downstream of a supraglottal constriction. Thus, if we wish to determine source characteristics from the speech signal, analysing the mixed-source blend is more elaborate than for single-source speech sounds. Moreover, as various authors have noted, the two sources are not entirely independent; in particular, the voicing source appears to modulate the noise source (Fant 1960; Flanagan 1972). Others have found that modulating the aspiration source during a vowel-to-voiced fricative transition leads to better-quality synthesis (Klatt and Klatt 1990; Scully 1990; Scully et al. 1992). While such interaction of sources inevitably complicates the model used for synthesis, and the analysis problem, it may also be the key to a more accurate model of the production mechanism itself. Closer study of the source interaction could lead directly to better quality synthesis of voiced fricatives and, potentially,

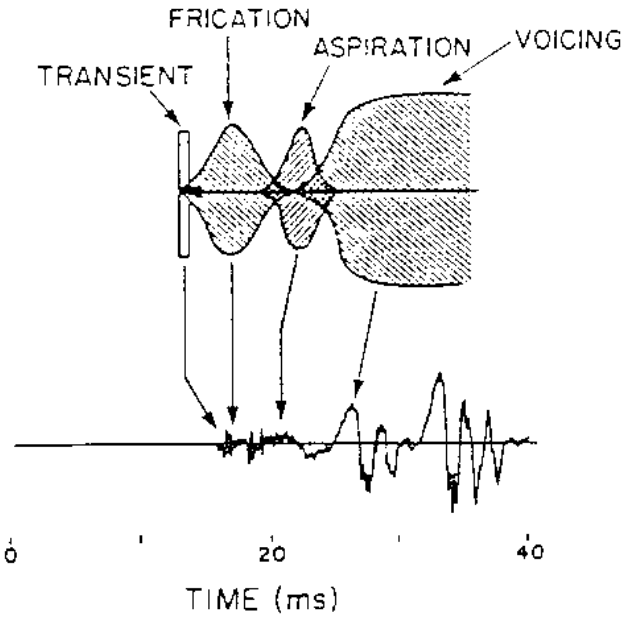


Figure 1.2: Effect of the different types of sound source during the release of an unvoiced, aspirated plosive (Stevens 1993, Fig. 6, p. 371).

of other mixed-source signals, such as breathy vowels.

1.2.2 Plosives

The interruption of the flow during the occlusion of a stop consonant causes a build-up of pressure in the oral cavity. When the obstruction is removed, the sudden release allows air to be expelled suddenly; the resulting sound is given the term *plosion*. For unvoiced plosives, e.g., /p, t, k/, the vocal folds are abducted to allow more air to flow through the tract; whereas for voiced plosives, e.g., /b, d, g/, an increase in subglottal pressure helps to create the conditions necessary for voicing, which usually has a lower glottal impedance associated with it (Titze and Story 1997). On release of a plosive, a whole sequence of sound production mechanisms is triggered. Initially, there is the burst; this is followed by frication as the obstruent articulator realises a narrow constriction at the point of approximation; then, for unvoiced stops in English, aspiration occurs before the onset of voicing (see Figure 1.2, Stevens 1993). However, even during voiced stops, the stop closure affects the air flow through the vocal folds and tract, and hence the production of turbulence noise. Research has shown (Fant 1973; Stevens 1993; Scully and Mair 1995) that a word-initial, aspirated, unvoiced stop, such as [p^h] in /pa/, follows the sequence:

- silence, while pressure builds up behind the point of closure;
- release, whose burst induces a transient response;

- frication, as there is a rapid flow through a small opening near the point of closure;
- aspiration, while there is considerable air-flow but no significant constriction in the vocal tract, as the vocal folds are being adducted;
- voicing, which begins once the vocal folds have been sufficiently adducted.

By classifying each of the stages from the time signal or spectrogram, their main features can be extracted from detailed examination, for instance, of the short-time spectra. Among others, we are interested in the aspiration stage and how the features adapt from normal speech to breathy and whispered modes of speaking. However, classification of the source types is not simple because they often overlap, and during voiced consonants they are also masked by phonation. To have a better idea of the term aspiration, let us refer to the literature to see how others have described the phenomenon.

1.2.3 Aspiration noise

Breath noise is present, to some extent, every time air flows through the vocal tract, which means that it accompanies almost everything we say. There are also complex patterns of events at the release of stop consonants and at the initiation and termination of voicing, not to mention interactions between phonation and the production of aspiration noise. Not surprisingly, there have been many contrasting (often rather informal) descriptions:

Aspiration is ...

"... big breath" (Dixit 1983);

"... a puff of air" (Ladefoged 1985);

"... the act of delaying the onset of voicing momentarily while exhaling air through a partially open glottis" (Deller et al. 1993);

"... essentially the same as frication noise, except that it is generated in the larynx." (Klatt 1980);

"... characterised by an *h-like* noise originating from a random source at the glottis or from a supra-glottal source at a relatively wide constriction exciting all formants." (Fant 1973).

In general, there is agreement that aspiration requires increased airflow, that it is manifested as a "noisy" signal, but that it is not frication. However, researchers disagree about the role of the glottis, the location of the source and the sense of cause and effect in relation to its properties. So, aspiration, which is generally accepted to be *h*-like noise produced by turbulence,

is often also associated with one or more of the following (depending in part on whether the description is given by a phonetician or a speech scientist):

- large airflow (Dixit 1983; Bristow 1984; Ladefoged 1985; Deller et al. 1993; Stevens 1993; Scully and Mair 1995),
- wider glottis configuration (Fant 1960; 1973), also Allen (1953) and Kim (1970) as cited in Dixit (1983),
- voicing lag (Lisker and Abramson 1964; Abercrombie 1967; Catford 1977; Ladefoged 1985; Deller et al. 1993; Scully and Mair 1995), and Ladefoged et al. (1976) as cited in Dixit (1983),
- glottal friction (Fant 1960; 1973; Kent and Read 1992; Deller et al. 1993; Stevens 1993; Scully and Mair 1995; Johnson 1997).

Yet, despite the disagreement concerning which feature of aspiration is the defining one, its main characteristic appears to be that it is accompanied by some form of broad-band, noisy sound source. Thus, we will consider it to be turbulence noise that is not caused by a constriction in the supralaryngeal vocal tract (i.e., not frication), and we will use the consensus of definitions to frame a working definition of aspiration: "flow-induced turbulence noise that is not frication".

Hence, in trying to discover the mechanisms by which aspiration is generated, we will consider such inextricable effects as the mode of vibration of the vocal folds (for voiced aspirates) and articulatory dynamics. Hoarse, breathy and whispered speech clearly contain increased amounts of aspiration, compared with normal, or modal, speech and therefore constitute a significant related area of study. Aspiration noise can also be partitioned into three principal classes: (i) constant flow [voiceless], (ii) steady harmonic [voiced], and (iii) transient [voiced/voiceless]. Accordingly, we aim to derive a generalised model through studying: (i) the response of the vocal tract to flow sources, (ii) the nature of the source signals, and (iii) sound-generating mechanisms.

1.3 Speech modelling

Since Fant's seminal work forty years ago (Fant 1960), the traditional approach to modelling the speech signal has been as the linear combination of an acoustic source (located either at the glottis, as in voicing, or elsewhere in the vocal tract) and a filter, representing the acoustic response of the vocal tract to the source. Fant calculated the vocal-tract transfer function from area functions, which he obtained by fusing X-ray profiles with cross-sectional tracings derived from knowledge of the changing cross-sectional profile along the vocal tract. The results of his predictions, which closely matched the properties of real speech, demonstrated how a good

approximation could be achieved using just an area function and a one-dimensional, plane-wave model. Since then, other measurement techniques have been used to gather more precise articulatory data, either from regions of special interest (e.g., near a supraglottal constriction by electropalatography, Stone 1991) or more generically for static configurations from medical imaging techniques, such as cathode-ray tomography (CT scanning) or MRI. Recently, these techniques have been improved to yield higher time resolution (Masaki et al. 1999; Mohammad 1999).

Incremental improvements to a model capturing these features can be readily incorporated into an articulatory synthesiser to generate more natural speech, but we also need an acoustic model to help investigate other aspects of the speech production system. An accurate and comprehensive model strengthens the links between the geometry and the actual sounds produced, and enables us to understand the relative importance of individual elements of the vocal tract, such as the sublingual cavity and the pyriform sinuses.

1.3.1 Filter models

The problems with a one-dimensional model occur when there is significant variation of the area function from this form, that is when cross dimensions are comparable to the lengths along the tube and when there are sizeable side branches. Real sound naturally propagates in three dimensions, and so considerable attention must be paid to how best one might try to accommodate a side branch into a plane-wave model (Dang et al. 1997). Also, the cross modes have a significant influence. For frequencies above the cut-on, which lies between 5 kHz and 7 kHz normally, acoustic modes exist out of the plane of propagation that was assumed, bringing extra poles and zeros into the VTTF; below the cut-on frequency, the modes are evanescent, i.e., non-propagating, and alter only the phase of the response. The phase adjustment can be incorporated into the model as a slight extension of the narrower tube elements in the overall area function. These extensions are referred to as end-corrections. Naturally, this step can be avoided using a model of higher dimensionality, for instance by finite element modelling (Motoki et al. 2000), with the consequent increase in complexity and an equal need for high-resolution geometrical details. While it is worth bearing in mind the existence of cross modes, most of the information transmitted in speech resides in the first 4 kHz or so (e.g., the formants and the lower harmonics of the fundamental frequency), as suggested by the intelligibility of telephone speech and the ear's perceptual weighting of frequencies. We are interested in a broader bandwidth than this for high quality speech, but it has been suggested that the bend in the vocal tract at the top of the pharynx attenuates the transmission of cross modes (Liljencrants 1985).

Still, as the filter becomes more realistic, its inverse can be applied to the acoustic signal to give a better description of the acoustic sources, which is one of our specific objectives for

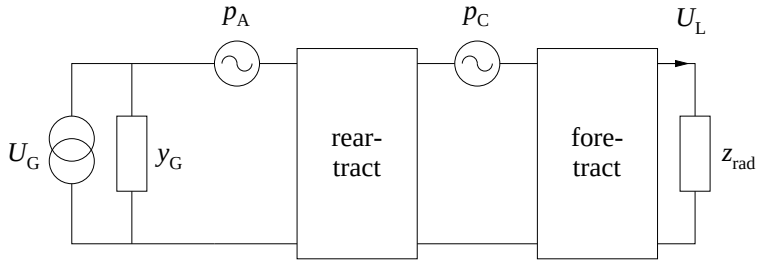


Figure 1.3: Vocal-tract model containing two glottal sources U_G and p_A, and a supraglottal source p_C.

unvoiced sounds. Yet, to explain any aero-acoustic system fully, fluid dynamics must ultimately be taken into account, which may involve flow separation, turbulent mixing and non-isentropic sound propagation. Flow separation and the formation of a jet are known to be extremely sensitive to local geometries, which fact increases the necessity of extracting high-quality cross-sectional areas and related geometrical data. So, there is a strong imperative to make fuller use of the MRI and other data that are becoming available to speech scientists to enhance the current performance of vocal-tract acoustic models.

For voiced speech, the filter can be considered as the acoustic response of the pharynx, oral cavity and nasal cavity from a source at the glottis as it is radiated to the far field, ignoring the coupling of the subglottal cavities (i.e., the trachea and lungs). The nasal cavity is decoupled when the velum is raised, which is true for all fricatives and plosives, and is generally true for the majority of British English vowels. In these cases, the airway is just the pharynx and oral cavity, and can be approximated by a single concatenation of short tube sections. Indeed, many studies that predicted the vocal-tract transfer function (VTTF) used straight, rigid tube sections and assumed axial, plane-wave propagation. The transmission of sound to the far field is typically treated as an ideal piston acting in either an infinite or a spherical baffle. Even though interactions exist between the source and the acoustic loading of the vocal tract, the insights derived from such simple predictions have gone a long way to capture the essential characteristics of the vocal-tract acoustics and to explain experimental observations.

1.3.2 Source models

In the search for accurate speech production models, acoustic studies abound where speech signals are dominated by a single source of sound, such as modal phonation or voiceless frication; for signals comprising contributions from a mixture of sources, the difficulty of separating them has hindered the interpretation of detailed analyses. Consequently, vowels and voiceless fricatives have received much attention, allowing for the development of functional source models (Fant 1960; Flanagan 1972; Ishizaka and Flanagan 1972; Shadle 1985, 1995a; Stevens 1971,

1998; Titze 1994), extensive parameterisation of corpora (Löfqvist et al. 1995; McGowan et al. 1995) and validation against a wide range of articulatory data, such as electro-glottography (EGG), electro-palatography, magnetic resonance imaging (MRI) and X-ray (Fant 1960; Badin 1991; Narayanan et al. 1995; Shadle and Scully 1995). Figure 1.3 gives a schematic representation of this combination of source and filter, showing three possible source configurations: U_G (with admittance y_G), p_A and p_C. The radiated sound is determined by the volume velocity at the lips U_L, which is shown as the current through the radiation impedance z_{rad} in this model. These models, with minor modifications, have provided a reasonable means of synthesising mixed-source sounds, like plosives and voiced fricatives, but such sounds remain, for the most part, under-explored and do not have commensurate, physically realistic models to relate the acoustic signals to the aero-acoustic phenomena that produce them. In this study, we have tried to take the first steps towards producing such a model by using non-invasive in vivo measurements, and analysing the radiated acoustic signals with a view to discovering their aero-acoustic characteristics.

In mixed-source speech, the characteristics of the individual sources are unclear; during analysis, one source's features are often obscured by the other sources, which increases the errors in estimating the source parameters. For example, the spectral tilt of the voicing source can be grossly underestimated in the presence of frication noise. This reduction in signal-to-error ratio (SER) of the signal parameters obscures the source properties, conceals patterns in the data and presents an obstacle to the determination of the cross-coupling between sources. As has been shown (Crow and Champagne 1971; Simcox and Hoglund 1971), the interaction of acoustic waves and turbulent flow can be very strong and highly non-linear, which further exacerbates the difficulties in identifying any source mechanism. Nevertheless, voiced fricatives, such as /v, ð, z, ʒ/, like aspiration, have received some attention as representatives of an essential phonological category. The characteristics of the voiced component of the sound are embodied fairly accurately by a typical source and filter, using one of the established vocal-fold models, e.g., Ishizaka and Flanagan (1972). The frication component, which dominates the unvoiced part of the voiced fricative, shares many features with its voiceless counterpart (i.e., /s/ for /z/). Many researchers have observed these common features: their spectra (in the high frequency region, Shadle and Scully 1995), transfer functions (both predicted and measured, Badin 1991), and articulatory configurations (Narayanan et al. 1995). Some spectral characteristics are altered, but the most conspicuous difference in frication between the voiced and unvoiced fricatives is the pulsing, during phonation, of the turbulence noise. Furthermore, the masking and assimilation effects are likely to augment the perceptual impact of these temporal differences on the listener (Hermes 1991).

In simple models of voiced fricatives, the voicing and frication sources are inserted into

the system and the output is formed from the sum of their individual contributions: voicing as a volume-velocity source at the glottis; frication as a pressure source at the supraglottal constriction. Although Fant (1960) noted that source-source interaction occurred as "periodic and synchronous" modulation of the frication source by phonation, Flanagan's electrical analogue model was one of the first to incorporate modulation of the fricative source amplitude (Flanagan and Cherry 1969). Band-passed Gaussian noise (0.5–4 kHz) was multiplied by the square of the volume velocity at the constriction exit U_n, which included the d.c. component, to give the pressure (voltage) source P_n in series with a variable source resistance R_n. Sondhi and Schroeter (1987) employed a similar model for a practical implementation of an aspiration source at the glottis, gated by a threshold Reynolds number; for frication they placed a volume-velocity source P_n/R_n one section (0.5 cm) downstream of the constriction exit (or at the lips for /f, v, θ, ð/), because of poor subjective results with pressure sources.

In Scully's work (Scully 1990; Scully et al. 1992), the source generation is based on Stevens' result from static experiments (Stevens 1971): the strength of the pressure source p_s is proportional to $\Delta P^{\frac{3}{2}}$, where ΔP is the pressure across the constriction. This source, depending on slowly-varying articulatory and aerodynamic parameters, was applied equally to aspiration and frication sources. Since ΔP across the supraglottal constriction is lower for voiced than voiceless fricatives, this equation partially accounts for the weaker frication source. These parameters do not encode any modulation, or allow for the flow separation lag in jet formation (Pelorson et al. 1997). However, motivated by the results of perceptual tests, the aspiration source was modulated using the rapidly-varying glottal area. Klatt, treating aspiration and frication identically, modulated the noise source with a square wave (50 % burst duration) that was switched on during voicing, remarking that it is "not necessary to vary the degree of amplitude modulation ..., but only to ensure that it is present" (Klatt 1980). In an analysis-by-synthesis procedure, Narayanan and Alwan (1996) used a combination of pressure (dipole) and volume velocity (monopole) sources to match measured fricative spectra, and concluded that the monopoles should be placed at the constriction exit and the dipoles at one or more obstacles: at the lips for /f, v, θ, ð/, at the teeth for /s, z/ and at the teeth and vocal-tract wall for /ʃ, ʒ/.

None of the above models considers any non-acoustic fluid motion, yet in a flow duct experiment (Coker et al. 1996), the arrival time of a pulse of radiated noise, depending strongly on the constriction-obstacle distance, suggested a convection velocity of less than half the flow velocity at the jet exit (8 m/s). In his recent PhD thesis, Sinder (1999) presents a model for fricative production that is based on aero-acoustic theory. Once the necessary flow-separation conditions have been met, vortices are shed, which convect along the tract, generating sound as they go, particularly when encountering an obstacle. Therefore, it is desirable to consider

both acoustic and aerodynamic mechanisms.

1.4 Speech analysis

Speech analysis describes the process of drawing out pertinent features from a recorded speech signal, and estimating any parameters associated with those features. For example, on the release of a stop consonant, the signal contains a sudden transient disturbance, whose time of incidence and peak amplitude might constitute an apt description. Indeed, many kinds of features have been identified by researchers as valuable indices for quantification, comparison and classification of speech tokens, and several classes of features have emerged for describing different aspects of different classes of phoneme. For instance, the incidence time would be irrelevant for the incremental growth of an unvoiced fricative, which may however have spectral features that are relevant for perception and recognition.

This section describes a variety of features, which relate to the amplitude, timing and spectral properties of the speech signal, and to phonation. Later, we introduce techniques for separating the voiced part of the signal from the remainder to ameliorate characterisation of each part individually.

1.4.1 Features of the speech signal

Naturally, amplitude features, such as the mean amplitude or the overall sound pressure level (SPL) during the steady portion of a phone, and the amplitude of peaks can be readily obtained from the time-series signal. The SPL is expressed on a logarithmic scale of decibels, or dB:

$$\mathrm{SPL} = 10 \log_{10} \frac{\langle p^2 \rangle}{p_{\mathrm{ref}}^2}, \tag{1.1}$$

where $\langle p^2 \rangle$ is the mean squared sound pressure (or mean intensity), which is related to the nominal level for the threshold of hearing, $p_{\mathrm{ref}} = 20\,\mu\mathrm{Pa}$. By passing the signal through an array of band-pass filters, similar quantities can be obtained for different frequency bands. Often it is more useful to consider relative levels than absolute ones, so the differences between two frequency bands, say, or how a parameter varies over time may be more descriptive of the underlying speech, and thus be a better feature.

Timing features are a critical factor for capturing aspects of the dynamics of speech, yet a gross measure of word rate, say, is of little value in detailed phonemic analysis. Any identified signal feature can be associated with the time when it occurred, but relative timing, as alluded to above, adds both insight and generality. For example, the time measured from the burst at the release of a preceding plosive to the onset of larynx excitation, the voice onset time, has been shown to hold perceptual salience for classification of plosives. Other examples include the duration of vowels and even the period between successive pitch pulses.

Voicing is such a prevalent and important part of speech that it has earned many measures of its own. Not only are the onset and offset times critical in relation to other speech features but, during periods of phonation, the oscillation of the vocal folds, its amplitude, its frequency and its regularity may all be quantified. The fundamental frequency of oscillation f_0, which is closely associated with pitch, is central to studies of prosody and pitch accent, and is often a vital first step for further analyses. During each glottal cycle in modal voicing, there is a period in which the folds are together and the glottis is essentially closed, and a period for which it is open. The ratios of these periods to the total pitch period, also known as the glottal cycle, are called the *closed quotient* and *open quotient* respectively, and are a useful measure of voice quality. They are particularly useful in the analysis of dysfunctional voices, where closure is typically incomplete and poorly defined, and of singing, since these parameters change dramatically in the training of a singer (Howard 1999). In pathological speech, voicing can be highly irregular and normal speech always contains some degree of perturbation from a smooth trajectory (Murry et al. 1979). Variations in pitch are known as *jitter* (Lieberman 1961) and variations in amplitude as *shimmer* (Koike 1969), both of which have been used as measures of hoarseness themselves (Hanson et al. 1997; Awan and Frenkel 1994), and are discussed in greater depth in Chapter 5. These are normally concomitant and highly correlated in speech, and tend to have a characteristic frequency of variation. The combined perturbations that are observed are termed *flutter* if they are rapid (of the order of 5 Hz); slow variations, over the length of a syllable, are termed *wow*, which is effectively a prosodic property. For example, Klatt and Klatt (1990) used parameters related to flutter and diplophony (shimmer at $f_0/2$) to improve the naturalness of their speech synthesiser.

Features of the spectrum have been an increasing area of interest and, with the advent of widely-available computer applications for calculating the Fourier transform, of the spectrogram (aka. sonogram or periodogram) and other spectral representations. For instance, the most striking property of a vowel spectrum (apart from the periodic striping at the harmonics of f_0 in narrow-band spectra) is the resonances that have been excited, evidenced as broad spectral peaks with bandwidths typically in the range 40–400 Hz. The amplitude, centre frequency and bandwidth of these resonances, or *formants*, describe the signal in a way that may be transposed into poles in a system representation. They are linked to the acoustic resonances of the vocal-tract transfer function. This view of the speech signal, as the product of a time-varying linear system is not only extremely useful for faithfully characterising salient features of the signal, but it demonstrates the power of model-based analysis over more traditional parameterisations. The method of inverse filtering (e.g., Rothenberg 1973) attempts to remove optimally the effects of the formants from the speech signal, so that what remains is seen as the source waveform. However, there are also anti-resonances, or zeros, in the VTTF which

appear as troughs or valleys in the speech spectrum for all classes of sound. These are the result of other branches in the tract, such as sinuses and the subglottal airways. In fricatives, which are excited by a localised supraglottal source, the part of the tract upstream from the constriction (the rear-tract) produces zeros, and for nasals, the oral cavity acts as a side branch to a similar effect.

The amplitude of peaks and troughs, including the height of the first couple of f_0 harmonics, have been variously combined to give useful spectral features, like spectral tilt, which have then provided the main data source for correlation and other studies (Stevens and Hanson 1995; Hanson 1997; Shadle and Mair 1996; Jesus and Shadle 2000). In fact, the levels of low and high frequency regions have been compared as an indication of the relative amplitude of voiced and unvoiced sources, which has come to be known as the harmonics-to-noise ratio.[1] Being of direct significance to those engaged with speech pathologies and synthesis alike, a number of techniques has been developed for HNR estimation (Yumoto et al. 1982; Muta et al. 1988; Cook 1991; de Krom 1993; Awan and Frenkel 1994; Michaelis et al. 1995; Qi and Hillman 1997; Qi et al. 1999; Murphy 1999). Also, there are many perceptual characteristics, such as roughness, breathiness and hoarseness, that are commonly used by speech clinicians. Although they can be strongly correlated to calculable signal attributes, like the HNR, they are not within the scope of the present study.

1.4.2 Decomposition techniques

The acoustic cues that are central to our ability to perceive and recognize speech derive from a variety of acoustic mechanisms and are often classified according to the nature of the sound source: voicing, frication, plosive or aspiration (Stevens 1993; Scully and Mair 1995). Identifying and characterising the various sources is fundamental to speech production research (Fant 1960; Flanagan 1972; Stevens 1998), and to the classification of pathological speech. Recent studies of hoarse speech have concentrated on measures of roughness in phonation, e.g., Herzel (1993), and yet turbulence-noise sources contribute largely to this effect (as breathiness). In normal or pathological speech, when more than one sound source is operating, it is difficult to segment the corresponding acoustic features, which typically overlap both in time and frequency, thus hindering the isolation of individual source mechanisms, and making it practically impossible to examine source interactions in any detail. Our particular area of interest is turbulence-noise sources in the vocal tract and, to explore these phenomena, we would like to be able to analyse the voiced and unvoiced components of mixed-source speech

[1] Sometimes an effort has been made to model the voiced component explicitly, in which case modelling errors, noise disturbance, jitter and shimmer all contribute to the unvoiced component. Thus, the voice quality metric, the HNR, which is intended solely as a measure of the strength of unvoiced sounds in relation to voicing, is misleadingly under-estimated because of these additional contributions.

separately, possibly even to distinguish between all the different contributions. To that end, we have developed a signal analysis technique for separating the part attributable to voicing from the simultaneous, unvoiced parts. Assessing the relative contribution of these two components as a harmonics-to-noise ratio has long been a useful tool in the laboratory and the clinic, but there has been growing interest in more complete descriptions of the voiced and unvoiced signal components. Recent development of decomposition algorithms has been fuelled by the demands of numerous speech applications: enhancement (Silva and Almeida 1990; Graf and Hubing 1993; Hardwick et al. 1993; Damper et al. 1995; Yoo and Lim 1995; Logan and Robinson 1997), modification (Laroche et al. 1993; Stylianou 1995; Richard and d'Alessandro 1997), coding (Serra and Smith 1990) and analysis (Cook 1991; Feder 1993).

Decomposition is generally achieved by first modelling the voiced component deterministically, since voicing tends to be the larger signal component, and then attributing the residue to the estimate of the unvoiced component. Concentrating the voiced component into a certain region of a transformed space improves estimation of the model's parameters. The extraction of energy concentrations in the signal is equivalent to the separation of deterministic and stochastic elements, which may be realised by a thresholding operation, as in Donoho (1993) using wavelets. Serra and Smith (1990) combined peak-picking and tracking to code the voiced (deterministic) part and fitted line segments to the residual noise spectrum. However, the regularity of vocal fold vibration can be used to define the region of concentration, and to design a comb filter that effectively averages successive pitch periods. The two main approaches are time domain (TD) and frequency domain (FD), although most contain elements of both.

In the TD methods, the comb filter is periodic with teeth aligned on the pitch pulses. The models typically assume that noise is added to pulsed excitation of a time-varying, linear filter. To adapt the spacing of the teeth of the comb filter in synchrony with variations in voicing, knowledge of the glottal pulse instants is required. There have been many TD realisations of this pitch-synchronous principle, which have accommodated timing variations by truncation and zero-padding (Frazier et al. 1976; Lim et al. 1978; Yumoto et al. 1982), scaling (Murphy 1999), least-squares alignment (Pinson 1963; Feder 1993) or dynamic time warping (Graf and Hubing 1993).

FD methods estimate the Fourier series of pitch harmonics from the short-time Fourier transform (STFT), using the fundamental frequency f_0 to identify regions of the spectrum that correspond to voicing. Thus, they model voicing by a short-time harmonic series (Parsons 1976; Griffin and Lim 1988; Silva and Almeida 1990; Laroche et al. 1993; Hardwick et al. 1993; Yoo and Lim 1995; Stylianou 1995), whose parameters tend to be smoothed between analysis frames. Griffin and Lim (1988) used the pitch harmonics to sub-divide the spectrum, and made a voiced/unvoiced decision on each harmonic band for coding the speech signal.

Technique	Area of speech research		
	Voice quality	Enhancement	Analysis/coding/modification
Comb filter	Yumoto et al. 1982; Awan and Frenkel 1994; Murphy 1999.	Shields Jr. 1970; Frazier et al. 1976; Lim et al. 1978.	Pinson 1963; Cook 1991.
Correlation-based	Michaelis et al. 1995; Qi et al. 1999.		
Cepstral	de Krom 1993; Darsinos et al. 1995; Qi and Hillman 1997.		d'Alessandro et al. 1995; d'Alessandro et al. 1998; Richard & d'Alessandro 1997; Yegnanarayana et al. 1998; Gabelman et al. 1998.
Asynchronous		Parsons 1976; Griffin and Lim 1984; Silva and Almeida 1990; Hardwick et al. 1993; Yoo and Lim 1995; Damper et al. 1995.	Serra and Smith 1990; Laroche et al. 1993; Feder 1993; Deller et al. 1993; Stylianou 1995.
Pitch-scaled	Muta et al. 1988.		Jackson and Shadle 1998, 2000c.
Dynamic time warping		Graf and Hubing 1993.	
Wavelet		Donoho 1993.	
AR-HMM		Logan and Robinson 1997.	

Table 1.1: Summary of literature relevant to decomposition of speech signals, where AR-HMM signifies auto-regressive hidden Markov modelling.

A compromise was proposed by de Krom (1993), who created a harmonic comb filter from the rahmonics of the real cepstrum (de Krom 1993; Darsinos et al. 1995; d'Alessandro et al. 1995; Qi and Hillman 1997; Yegnanarayana et al. 1998). The log-spectrum thus obtained from the harmonic cepstrum (with the spectral envelope now removed), which oscillates about zero, was thresholded: frequencies for which it was greater than zero were defined as harmonic, and those less than zero as anharmonic. Hence, the partitioning of regions in the cepstral domain provided a means of labelling those regions in the STFT spectrum. Table 1.1 contains a summary of the relevant literature, briefly indicating each article's main area of application and the basis of its method.

Still, choosing a technique for one's own data and purpose is not straightforward. Lim, Oppenheim, and Braida (1978) showed that TD comb filtering decreased intelligibility, whereas a harmonic method increased it (Hardwick, Yoo, and Lim 1993). On the other hand, Qi and Hillman (1997) found that an adaptation of de Krom's method performed poorly compared to a TD method (Yumoto et al. 1982). Although some techniques effectively applied a rectangular window, most have chosen a smooth function, i.e., Hann or Hamming. All of these FD methods use a frame of fixed duration, so that the spacing of the pitch harmonics is proportional to f_0, which implies that they do not generally coincide with the STFT bins. The leakage and smearing caused can be accounted for (Silva and Almeida 1990), but the concentration of the harmonics can be significantly improved by forcing the spacing to coincide with the frequencies of the STFT bins, which is achieved with an integer number of pitch periods in the time frame. Computing the STFT pitch-synchronously again requires knowledge of the pulse instants (Murphy 1999), but scaling the frame size to the local estimate of the pitch period avoids this drawback, which is the approach that we have taken with the pitch-scaled harmonic filter (PSHF).

For HNR estimation and synthesis applications (coding, copy-synthesis, modification), the accuracy with which the true component is estimated is not important provided the salient signal properties are captured, which is also the case for certain types of analysis. More generally, though, we would like to analyse all the available information and nothing else, and therefore to provide an output with a minimum of distortion. After subtraction of the voicing model from the original spectrum, the residue's spectrum typically lacks data at the harmonics, i.e., the region where voicing was concentrated, and values of zero may be the best estimate available for the unvoiced signal component. Yet, for feature extraction from the power spectrum (or for generating a stochastic model that reproduces the longer-term spectral characteristics of the unvoiced component), interpolation can be advantageous. Interpolation has been done, for example, by linear prediction (Laroche et al. 1993) or by approximating the spectral envelope with line segments (Serra and Smith 1990) or cepstral coefficients (Stylianou

1995). One recently-published technique (Yegnanarayana et al. 1998) uses a reconstruction algorithm, but we have discovered certain problems with it, which are described in Appendix D.

1.5 Organisation of the thesis

To match the three main aspects of a general understanding of the acoustic theory of unvoiced speech production, a three-pronged approach has been used. First, work was continued on an existing vocal-tract acoustics program (VOAC) with a view to testing the scope of its functionality and fixing several of its faults. Aero-acoustic experiments using physical, flow-duct models were conducted as part of an earlier study (Shadle 1985), which were designed to reproduce the acoustic and flow properties of the human speech production system, and hence enable investigation of source mechanisms. These measurements were used to validate the predictions of the modified program, as described in Chapter 2.

Second, in Chapter 3, outlines of the vocal tract were generated from our library of magnetic resonance imaging data, which were interpreted to produce a description of the vocal-tract geometry that VOAC could use to predict vocal-tract transfer functions. The transfer functions gave insight to the spectral features of the speech signals, and a basis for comparison against recordings of the same subject. They were also used to synthesise phones from the respective images.

Third, a signal analysis toolkit has been assembled for examining speech recordings (primarily of sound pressure) and extracting information about plosive, fricative and aspiration-noise signals. It contains the short-time Fourier transform, auto- and cross-correlation functions, the spectrogram, time- and ensemble-averaging, linear prediction coding analysis and synthesis, and the cepstrum, which are discussed in Chapter 4, where the results are presented for plosives. To observe unvoiced sounds in the presence of voicing, a special tool, called the pitch-scaled harmonic filter (PSHF), has been developed that decomposes the speech signal into harmonic and anharmonic components.

The PSHF, presented in Chapter 5, provides outputs that constitute our best estimate of the voiced and unvoiced signals (suitable for time domain analysis), and spectrally-interpolated outputs that provide a better estimate of the components' power spectrum (suitable for power spectral analysis and modeling). Previous techniques have failed to distinguish these two objectives of the decomposition task. The performance of the PSHF algorithm was tested using synthetic speech signals which contained three kinds of disturbance: shimmer (perturbed amplitude), jitter (perturbed fundamental frequency f_0), and additive Gaussian noise with variable burst duration. Chapter 6 presents examples of the separation of a recorded speech signal into its periodic and aperiodic components including fricatives, pressed and breathy

vowels, and nonsense words.

In Chapter 7, the PSHF was used to open up the path for a new kind of analysis, which capitalises on the fact that the voiced and unvoiced output signals were produced simultaneously by the original speaker. By performing this kind of mixed-source analysis on voiced fricatives, we were able to investigate timing differences that led us towards a theory of modulation of the frication noise in voiced fricatives. Chapter 8 draws together the three strands of this study, summarises its main findings for plosives, fricatives and aspiration, and suggests potential routes for profitable research in the future. The appendices provide supporting details: Appendix A, the aero-acoustic equations; Appendix B, VOAC's implementation; Appendix C, vocal-tract anatomy; Appendix D, the periodic-aperiodic decomposition algorithm of Yegnanarayana et al. (1998).

1.6 Contributions

A number of publications has resulted from the research carried out for this thesis. They are either papers in peer-reviewed academic journals or contributions presented at international conferences, as listed below.

1.6.1 Journal articles

Jackson, P.J.B. and C.H. Shadle (2000b). Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *Journal of the Acoustical Society of America*, 108(4): 1421–1434, October 2000.

Jackson, P.J.B. and C.H. Shadle. Decomposing speech signals into their simultaneous voiced and unvoiced components. *IEEE Transactions on Speech and Audio Processing*, submitted April 1999, revised and re-submitted March 2000.

1.6.2 Refereed conference papers

Jackson, P.J.B. and C.H. Shadle (1998). Pitch-synchronous decomposition of mixed-source speech signals. In *Proceedings of the joint International Congress on Acoustics and Meeting of the Acoustical Society of America*, Seattle, WA, 1:263–264, June 1998.

Jackson, P.J.B. and C.H. Shadle (1999a). Analysis of mixed-source speech sounds: aspiration, voiced fricatives and breathiness. In *Proceedings of the 2nd International Conference on Voice Physiology and Biomechanics*, Berlin, Germany, p. 30 (abstract only), March 1999.

Jackson, P.J.B. and C.H. Shadle (1999c). Modelling vocal-tract acoustics validated by flow experiments. *Journal of the Acoustical Society of America*, Presented at the joint Meeting

of the Acoustical Society of America and European Association of Acoustics, Berlin, Germany, 105(2, Pt. 2):1161 (abstract only), March 1999.

Shadle, C.H., M.A.S. Mohammad, J.N. Carter and P.J.B. Jackson (1999). Dynamic magnetic resonance imaging: new tools for speech research. In *Proceedings of the International Congress on Phonetic Sciences*, San Francisco, CA, 1:623–626, August 1999.

Jackson, P.J.B. and C.H. Shadle (2000a). Aero-acoustic modelling of voiced and unvoiced fricatives based on MRI data. In *Proceedings of the 5th Speech Production Seminar*, Seeon, Germany, pp. 185–188, May 2000.

Jackson, P.J.B. and C.H. Shadle (2000c). Performance of the pitch-scaled harmonic filter and applications in speech analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 3:1311–1314, June 2000.

Chapter 2

Acoustic flow-duct modelling of the vocal tract

2.1 Overview

Speech generally involves the flow of air through the vocal tract which leads to the generation of sound. These sound sources are filtered by the tract, but also tend to incur flow-related losses. Classical programs ignore flow-related losses or lump them together with others. These losses are separately catered for in our **vo**cal-tract **ac**oustics program (VOAC), which is used to compute vocal-tract transfer functions (VTTFs) within the framework of the source-filter paradigm.

According to this kind of model, speech sounds can be thought of as the result of filtering an acoustic source signal. The filter is typically determined by the geometry of the supralaryngeal airways, referred to as the vocal tract, and the location of the source therein, as well as the effects of sound radiation from the lips to the far field. For voicing, the source is created through the interplay between the vocal-fold mechanics and the aerodynamics of the glottal air flow. For frication and aspiration, sound is generated from flow turbulence, mainly as it impinges on obstacles within the tract. For plosives, the burst noise is produced by the sudden release of pressure originating from the point of constriction, but components of other sources, such as frication and aspiration, are usually concomitant because of the flow. The source-filter model implicitly assumes that the sound source and the acoustic response of the filter are independent, neglecting any non-linear or time-varying interactions. However, one can have a source model and a filter that are not linear and time-invariant and, moreover, source dependencies can be included in the VTTF (and conversely effects of the filter into the source). The assumptions of independence classically made in speech research though are good to a first approximation, and can achieve synthesis of a reasonable quality. In addition, such

classical models have successfully been used to replicate and predict many observable features of real speech, most notably the formant frequencies.

The acoustic filtering of the vocal tract is represented in the time domain as an impulse response function, which can be convolved with any source signal to produce the predicted output at the lips. In the frequency domain, the ideal input-output behaviour may be described by a transfer function, which can be evaluated at any chosen frequency. In reality, the response is estimated from limited measurements, which are usually corrupted by noise and made at a set of discrete frequencies. These measurements are used to estimate the frequency response function (FRF). VOAC was designed to compute, as its principal output, the frequency response of the VTTFs for comparison with the measured FRFs.

When modelling sources other than voicing of the vocal folds, one must be able to place a source elsewhere in the tract. During phonation, the sound source at the glottis (usually represented by a volume-velocity waveform) excites the resonances of the entire vocal tract, which is downstream of the source, and there is little effect from the upstream airways, which are often treated as anechoic. In contrast, a frication source, for example, which may be represented by a pressure source downstream of the supraglottal constriction, excites the part of the vocal tract upstream of the source, the rear-tract, as well as the anterior part, the fore-tract. Although the tract as a whole continues to govern the acoustic resonances of the filter, the resonances of the rear-tract produce anti-resonances in the overall response, which are manifested as troughs in the frequency response of the VTTF.

The power of these predictions depends very strongly on the accuracy of the geometrical data that are used to compute them. The problems involved in acquiring precise vocal-tract dimensions during speech have limited the extent to which predictions can be compared to measured responses. While improvements have been made to medical imaging techniques, as we will later acknowledge, a more reliable test of the acoustic model employs practical experiments on physical models, which can be designed to mimic various aspects of the vocal tract. One important aspect on which we have focused is the effect of a net flow of air through the tract. Thus, towards the end of this chapter, we show how VOAC has been validated against experimental measurements of flow noise in a test rig, before comparing the predictions for realistic tract shapes to analysed speech recordings. In the next chapter (Chapter 3), VOAC was applied to some of our own tract measurements.

This chapter describes the work that has been done on translation, revision and enhancement of VOAC, which was tested against experimental measurements and compared with other results in the literature. It covers the basic principles of operation: discretisation of the vocal tract, acoustic transfer between elements, and the VTTFs produced as output. Details of derivation of the transfer equations are presented in Appendix A and a pseudo-code tran-

scription of the program is given in Appendix B. Extensive tests have been performed on the program to eradicate any bugs from the code and to verify its predictions using primitive geometries, and they are described in Appendix B.1. Results of the application of VOAC are given towards the end of this chapter and the next (Chapters 2–3), and proposals for future developments in Chapter 8.

2.2 Vocal-tract acoustics program (VOAC)

This section gives an overview of VOAC's history and its revision. Apart from development of the input and output inferfaces, the ability to include an acoustic source at locations other than the glottis is described, which yielded a significant extension to VOAC's functionality.

2.2.1 Background

VOAC was originally developed by Davies, McGowan and Shadle from a flow-duct acoustics program, designed for automotive exhaust systems (Davies 1988). The main differences in functionality over conventional duct formulations were the inclusion of recessed duct elements, such as may be found in a car's silencer, and flow. For modelling the acoustics of the vocal tract, where the complex anatomy contains side-branches and sinuses, and where air normally flows during speech, these are clearly advantageous. Moreover, to accommodate some of the more gradual area changes along the tract, elements with linear and quadratic area profiles were added. These elements, referred to as the RAMP and the CONE respectively, allowed sound to propagate in non-planar fashion (the RAMP with cylindrical wave-fronts; the CONE spherical ones), and could be used interchangeably with the other plane-wave elements. For parts of the vocal tract with a circular cross-section, losses from surface absorption are minimal. Because cross-sections of the vocal tract are not usually circular, the model includes the hydraulic radius r_h so that tract elements with greater surface area can have losses correctly related to surface area:

$$r_h = \frac{2S}{l_\pi}\,, \tag{2.1}$$

where S is the cross-sectional area and l_π is the perimeter.

The program was first presented at the Vocal Fold Physiology meeting in Denver in 1991 (Davies et al. 1993). The paper described the applications of VOAC to some previously published area functions (Baer et al. 1991; Fant 1960) and compared the reported formant measurements to the predicted resonance frequencies, with and without the effects of flow. The findings were promising, although as a practical tool for speech analysis, VOAC was still in its infancy.

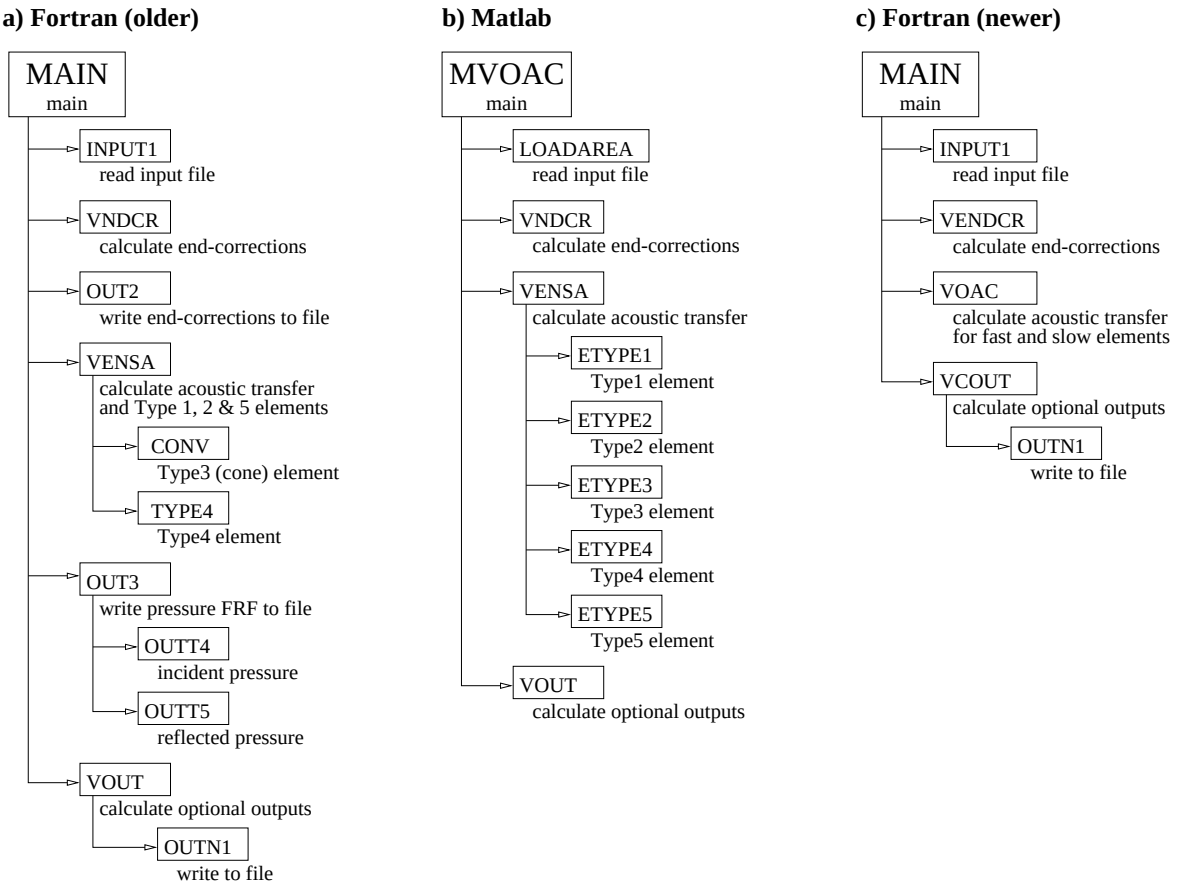


Figure 2.1: Program structures for different versions of VOAC: (a) v4.0 (older) written for Fortran 77, (b) v5.1.0 for Matlab 4.2, and (c) v4.5 (newer) for Fortran 77.

2.2.2 Translation into Matlab

At the start of the current project, the VOAC program consisted of a suite of Fortran files. After the detection of a few minor (mostly typographical) errors, a version of VOAC was successfully compiled. It was intended from the outset that the program be translated for use in a more flexible and graphic computer environment, and so this version of the program was held as a standard with which any later derivative could be made to comply. There were two essential reasons for translating the code: it needed to be (i) verified, which required the visibility of internal variables, and (ii) easily maintainable and upgradable for correcting program bugs and for extending the functionality. Matlab was chosen as the target language, since it met these criteria and for a number of subsidiary reasons: backwards compatibility with input files; modular program structure; high level control of mathematical floating-point operations; versatile graphical output; language in which the author was competent and for which licences were available; highly portable implementation.

The revision presented an opportunity to carry out minor alterations to the program structure, shown in Figure 2.1a. Motivated by simplicity and consistency, separate modules were

provided for each element type, which resulted in a more uniform and transparent structure, as shown in Figure 2.1b. Another version of the Fortran code later became available (shown in Figure 2.1c for comparison) that provided a valuable source of reference, but was not otherwise employed. The majority of the translation has been literal, avoiding any significant algorithmic changes. The original (Fortran) code contained many temporary internal variables, which reduced the complexity of mathematical expressions, but in Matlab these variables tend to clutter the workspace and some of them have been eliminated accordingly. The numerical accuracy of the program was improved in translation since floating-point values were stored as type `float` in the Fortran code (i.e., single float, 32 bits (4 bytes) or 9 significant figures in v4.0), and as `double` in Matlab (i.e., double float, 64 bits (8 bytes) or 19 significant figures in v5.1.0).

In the process of translation, the new code was repeatedly tested, to ensure that the output from the two programs was identical (to within the original numerical accuracy). However, testing with simple tube models, designed to call each module of the code, highlighted additional problems with this implementation (v5.1.0). Subsequently, a deeper examination of the underlying mathematics has been undertaken to verify the correspondence between the algorithm and the governing physical equations (see Appendices A and B).

2.2.3 Input

The input files for VOAC contain geometrical parameters, such as the area and hydraulic radius functions; aero-acoustic parameters, such as the net volume-flow rate and the speed of sound; and some arguments to control the program output, such as the number and range of frequencies at which the response is to be calculated. The geometrical description of the vocal tract, as area and hydraulic radius functions, is divided into elements, whose type is chosen from five possibilities: ORIFICE, RAMP, CONE, OUTLET and PIPE (described in Section 2.4.1). Accordingly, the input file states the number of elements, their types and their dimensions, followed by the remaining aero-acoustic and output parameters. Two highly desirable utilities for manipulating the area and hydraulic radius functions, to which we will refer jointly as *geometry functions*, display the functions graphically and generate them from other data sources, such as MRI. Although side-branches and curvature of the vocal tract can present minor challenges, the task of plotting the geometry functions with respect to distance along the vocal tract is technically trivial; interpretation of medical images, on the other hand, is less so and will be addressed in Chapter 3.

2.2.4 Output

In VOAC, sound waves are represented as sinusoidal partial pressures, $p^+(x,t)$ and $p^-(x,t)$, travelling over time t in positive and negative x-directions along the vocal-tract centreline

respectively. The program computes the complex amplitude (i.e., magnitude and phase) of p^+ and p^- at the end of each element, beginning from the lips. The pressure components resulting from an incident wave at the lips can be combined to calculate the acoustic pressure, velocity, force and impedance at the glottis. It is far more useful, however, to calculate the complete transfer function (TF) from the source location to the lips, where the sound is radiated to the far field. The acoustic loading of the radiated sound is accounted for by the inclusion of a radiation impedance.[1] Suitable combinations of the pressure components give the TFs from a pressure or volume-velocity source, $H^P(\omega)$ or $H^V(\omega)$ respectively, which have been added to VOAC to become its primary form of output:

$$H^V_{GL}(\omega) = \frac{U_L(\omega)}{U_G(\omega)}; \tag{2.2}$$

$$H^P_{GL}(\omega) = \frac{U_L(\omega)}{p_G(\omega)}, \tag{2.3}$$

where U is the volume velocity, p the acoustic pressure, and the subscript L refers to the lips and G to the glottis. The VTTFs can be further projected to predict the response at a point on the vocal-tract axis in the far field, using the expression (Shadle 1985, p. 102):

$$p_{\mathrm{ff}} = \frac{\omega\rho}{2\pi r} U_L(\omega)\,, \tag{2.4}$$

where ρ is the density of air, r is the distance to the point and $U_L(\omega)$ is the volume velocity at the lips, as a function of angular frequency $\omega = 2\pi f$. Hence, p_{ff} is what would be measured by a microphone at 0° and a distance r from the lips. Treating the complete response as the filter, it can be converted to a causal impulse response function in the time domain for convolution with a source signal, e.g., a train of glottal pulses for voicing, to calculate the acoustic signal at the microphone. (Some examples of synthetic speech sounds generated in this way will be described in Chapter 3.) There are new options to display TFs and the normalised radiation impedance, which augment those that were formerly available: glottal reflection coefficient, driving-point impedance at the glottis, attenuation of the incident wave, and wall impedance.

2.2.5 Intermediate source

A critical extension to VOAC allows acoustic sources to be placed at locations other than one end of the tract, as for the glottal source. The capability of driving the vocal tract with one or more intermediate sources, rather than just a terminal source, facilitates the modelling of the entire repertoire of human speech sounds, including plosives, fricatives and aspiration noise, and it offers the possibility of later including the subglottal airways.

Figure 2.2 is a simplified scheme showing an intermediate pressure source p_Q exciting the vocal tract, which is divided into the parts upstream and downstream of the source location,

[1] Changes were also made to the expression for the radiation impedance, which defines the boundary conditions at the lips (see Section 2.3.4).

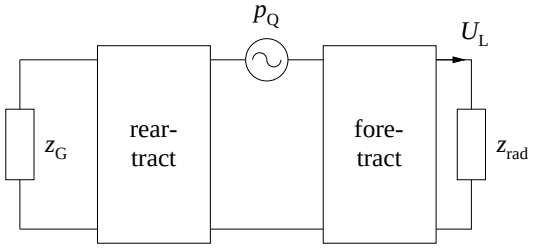


Figure 2.2: Diagram of an intermediate or supraglottal source in the vocal tract.

referred to as the rear-tract and fore-tract respectively. The acoustic load of the glottis is depicted by the glottal impedance z_G, and the radiation impedance z_{rad} as a load at the lips, where the volume velocity U_L exits. The implementation of the intermediate pressure source, which also requires modifications to the boundary conditions at the source location, is described in more depth in Section 2.4.2.

2.3 Acoustic formulation

This section discusses the acoustic formulation used within VOAC, which is based on wave propagation along a single axis. It provides a powerful framework for flow-duct modelling and can accommodate evanescent effects of cross modes, side branches and the radiation of sound from the open end.

2.3.1 Assumptions

Many simplifications need to be made in order to build a practical model of the vocal-tract acoustics, because there are many physical parameters for which it is difficult to obtain precise measurements in vivo, since they vary with time and with distance along the vocal tract, and they can be coupled, non-linear or simply unknown. As a practical measure, therefore, several assumptions have traditionally been made about the vocal tract, the air inside it and the sound waves that travel down it. A popular model is the classical electrical analogue (CEA, Flanagan and Cherry 1969), which assumes that:

the fluid is	... frictionless,
	... homogeneous, and
	... at rest, (i.e., no net flow);
the sound waves	... have small amplitude,
	... are axially propagating,
	... are isentropic,
	... have planar wavefronts;

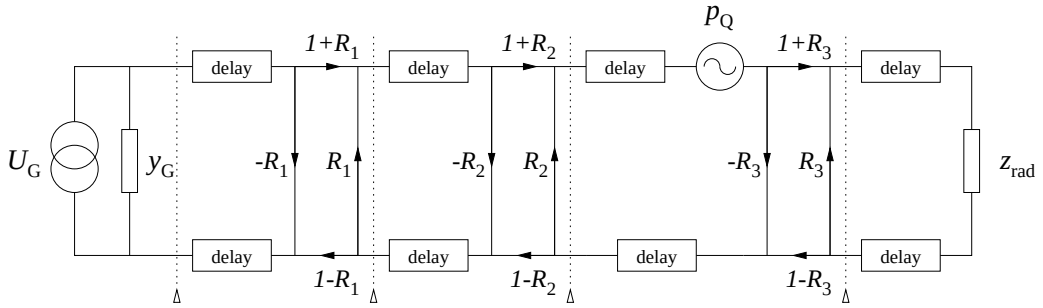


Figure 2.3: Transmission-line model diagram.

the wave-guide is a
- ... straight,
- ... rigid,
- ... static tube,
- ... with abrupt changes in shape along its length,
- ... radiating sound only from the open end.

Many of these assumptions have been relaxed, in some shape or form, in VOAC's calculations (Davies et al. 1993), but one that has not concerns radiation of sound from the mouth alone. For instance, sound may be radiated from the nasal port or through the walls of the vocal-tract, i.e., via the cheeks (Scully and Mair 1995), both of which have been neglected here. In the current implementation of VOAC and throughout this study, the influence of the nasal cavities has been ignored, because they are of lesser importance than the oral cavity for non-nasalised sounds, and because they do not change shape or their acoustic response. Nasalisation is a minor effect in British English vowels and all but irrelevant for fricatives, for which the velum must be raised to force air to flow through the supraglottal constriction. During nasals, liquids and the obstruent (closed) portion of voiced plosives, the nasal cavity plays an integral role, but these cases are not considered within the scope of this thesis.

2.3.2 Plane-wave basis

A plane-wave model is attractive because it offers a straightforward way to incorporate knowledge of the cross-sectional area into the model, as compliance and inertance components. Shown as a transmission line in Figure 2.3, the model consists of a volume-velocity source at the glottis, which is represented as an ideal current source (left) with the glottal admittance y_G in parallel, a sequence of concatenated tube sections, equivalent to delay elements, and finally terminated (on the right-hand side) by the radiation impedance z_{rad}. At the junctions between each section i, the transmission and reflection of the acoustic velocity (i.e., current) is determined by the reflection coefficients R_i. An intermediate pressure source p_Q is also shown, whose rear-tract is that part of the vocal tract to the left. In this model, electrical current is analogous to

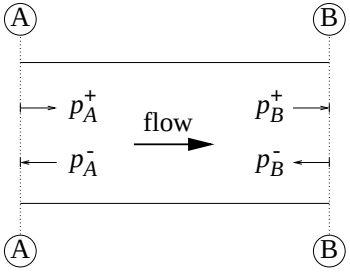


Figure 2.4: Plane-wave pressure components.

volume velocity and voltage, or potential difference, to pressure. Thus, the current through the radiation impedance U_L can be used to project the acoustics from the lips into the far field, as in Eq. 2.4. The transfer function can be calculated from the terminals of the network without any source at the glottis (i.e., an open circuit) for any on-zero frequency. Used in this way, the CEA provides a means of converting from area functions to TFs. The CEA is a direct analogy of a simple acoustic representation of the tract, which draws on the positive- and negative-travelling component pressure waves, as illustrated in Figure 2.4. The total sound pressure is represented by the superposition of the two pressure components, $p^+(x,t)$ and $p^-(x,t)$.

Using a Fourier approach, each of the two pressure components is described as a further superposition of a set of sinusoids which can be represented as the real parts of complex exponentials, with a corresponding magnitude and phase. Thus, it can be seen that the pressures at each end (A and B) of a rigid, uniform, lossless tube are related:

$$p_B^+(\omega)\exp(j\omega t) \;=\; p_A^+(\omega)\exp j\omega\left(t-\frac{l}{c_0}\right)\,, \tag{2.5}$$

$$p_B^-(\omega)\exp(j\omega t) \;=\; p_A^-(\omega)\exp j\omega\left(t+\frac{l}{c_0}\right)\,, \tag{2.6}$$

where l is the distance from A to B and c_0 is the speed of sound. Hence, the complex amplitude $p^+(\omega)$ becomes shorthand for $p^+(\omega)\exp(j\omega t)$ at any angular frequency ω and time t, and similarly $p^-(\omega)$ for the wave travelling upstream.

The total acoustic pressure at any point A, being the sum of these components, can be written as $p_A = p_A^+ + p_A^-$, and so Eqs. 2.5 and 2.6 can be combined at any particular time:

$$p_B(\omega) = p_A^+(\omega)\exp\left(-j\frac{\omega l}{c_0}\right) \;+\; p_A^-(\omega)\exp\left(+j\frac{\omega l}{c_0}\right)\,. \tag{2.7}$$

For plane waves, the acoustic velocity u_A^+ equals $p_A^+/\rho_0 c_0$, where ρ_0 is the density of air, and so the total acoustic velocity is $u_A = (p_A^+ - p_A^-)/\rho_0 c_0$. Hence,

$$u_B(\omega) = \frac{1}{\rho_0 c_0}\left[p_A^+(\omega)\exp\left(-j\frac{\omega l}{c_0}\right) \;-\; p_A^-(\omega)\exp\left(+j\frac{\omega l}{c_0}\right)\right]\,. \tag{2.8}$$

These conventions are followed in Appendix A, where the relations are further developed to express the transfers at abrupt area changes including the effects of net air flow.

2.3.3 Transfer at an abrupt area change

Ideal acoustic propagation occurs as an adiabatic and isentropic process, which implies that both mass and momentum are conserved for any control volume. At an abrupt area change, we can use the acoustic descriptions of these conservation laws to relate the pressures on either side: $S_B u_B = S_C u_C$ (mass); $p_B = p_C$ (momentum), where S_B and S_C are the cross-sectional areas to the left and to the right of the junction respectively as shown in Figure 2.5. Solving these for p_C gives us the paired equations:

$$p_C^+ = \left(\frac{S_C + S_B}{2S_C}\right) p_B^+ + \left(\frac{S_C - S_B}{2S_C}\right) p_B^-, \tag{2.9}$$

$$p_C^- = \left(\frac{S_C - S_B}{2S_C}\right) p_B^+ + \left(\frac{S_C + S_B}{2S_C}\right) p_B^-. \tag{2.10}$$

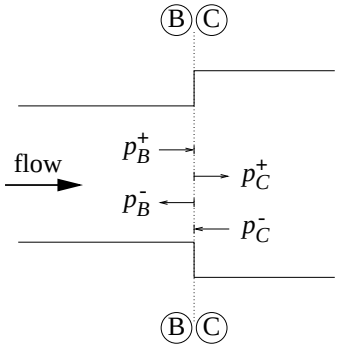


Figure 2.5: A simple expansion geometry.

Once any pair of pressures p^+ and p^- has been specified, the magnitude and phase of the pressures at any other point in the duct network can be calculated iteratively by applying these two principles: Eqs. 2.5 and 2.6, and Eqs. 2.9 and 2.10. Thus, if we know the pressures at the glottis, we can predict the outcome at the lips. Conversely, if we impose boundary conditions at the lips, we can deduce the response at the source and, assuming linearity and reciprocity, calculate the TF between these two locations. So how can we define the relative magnitude and phase of the incident and reflected pressures at the lips?

2.3.4 Radiation impedance

Depending on the wavelengths of interest, differing assumptions may be made so as to use standard results to calculate the radiation of sound from the oscillating sound field at the lips. For lower frequencies ($f < 1\,\text{kHz}$), the head is small relative to the wavelength and the vocal tract may be approximated by a semi-infinite tube radiating into free space. For higher frequencies ($f > 1\,\text{kHz}$), the head is large relative to the wavelength and the acoustic field at the mouth may be approximated by a piston in a sphere or in an infinite baffle. The latter, which is the version that is usually employed for speech, is defined as (Beranek 1954; Morse

1981; Kinsler et al. 1982):

$$z_{\text{rad}} = \pi a^2 \rho_0 c_0 \left(1 - \frac{2}{\omega} J_1(\omega) - jX(\omega)\right), \tag{2.11}$$

where

$$\begin{aligned} X(\omega) &= \frac{4}{\pi} \int_0^{\frac{\pi}{2}} \sin(\omega \cos\theta) \sin^2\theta \, d\theta \\ &= \frac{4}{\pi} \left(\frac{\omega}{3} - \frac{\omega^3}{3^2.5} + \frac{\omega^5}{3^2.5^2.7} + \cdots\right), \end{aligned}$$

and J_1 is a first order Bessel function of the first kind, $\omega = 2ka$ is the normalised angular frequency for the wave number, $k = \omega/c_0$, speed of sound c_0, and radius of the piston a. The radiation impedance is a way of relating the acoustic pressure to the acoustic velocity at the lips, as a consequence of the radiation from the mouth, which acts as an acoustic load. An alternative formulation equivalently relates the pressure components by means of a reflection coefficient, $R = p^-/p^+$, which is linked to the radiation impedance by the expression:

$$z_{\text{rad}} = \rho_0 c_0 \frac{1+R}{1-R}. \tag{2.12}$$

Alternative expressions for z_{rad}, some with flow, can be found in Davies et al. (1980), Davies 1988) and Munjal (1987). Note that VOAC does not include the glottal admittance (i.e., $y_G = 0$), which affects the way the acoustic source is transmitted into the vocal-tract filter. For glottal sources, it is actually a strong function of the glottal area $A_G(t)$. In such cases, a simpler approach is often to compensate the source function by modifying it before insertion into the filter model.

2.3.5 Cross modes

The implicit assumption when using a plane-wave model is that the effects of other modes of propagation are negligible. These modes account for the matching of the acoustic fields in three dimensions at a spatial discontinuity, such as an abrupt area change. However, the non-planar oscillations are generally evanescent and are unable to support the propagation of any radiating sound for frequencies below the first cross mode, which occurs at the cut-on frequency $f_{\text{cut-on}}$. They, therefore, do not contribute to the net acoustic response of the vocal tract for $f < f_{\text{cut-on}}$, except reactively. Their effect can be modelled as an adjustment to the effective length of the tube section, known as an end correction (Morse and Ingard 1968). The cut-on frequency places an upper bound on the range of frequencies for which a plane-wave model is wholly valid. Above the cut-on frequency, while cross modes are capable of propagating acoustic energy, they are not likely to be as strongly excited as any axial modes, but predicted acoustic responses should nevertheless be interpreted with due caution.

With a circular cross-section, the standard values for the cut-on frequencies of the first three cross modes are $(kr)_1 = 1.84$, $(kr)_2 = 3.0$, and $(kr)_3 = 3.8$ for a radius r and wave number k

(e.g., Davies 1988, p.92). In the presence of flow the first one becomes $(kr)_1 = 1.84\,(1-M^2)^{1/2}$. For the duct models to be discussed in Section 2.5.1, the cut-on frequency is:

$$\begin{aligned} f_{\text{cut-on}} &= \frac{1.84\,(1-M^2)^{\frac{1}{2}}\,c_0}{2\pi r_{\max}} \\ &\approx 7.95\,\text{kHz} \qquad\qquad (r_{\max} = 1.27\,\text{cm},\ c_0 = 344.8\,\text{m/s}), \end{aligned} \tag{2.13}$$

over the entire range of flow rates: $0 \le U \le 420\,\text{cm}^3/\text{s}$. For the human vocal tract in a typical vowel configuration, we obtain a lower value (with $r_{\max} = 2.0\,\text{cm}$, $c_0 = 359\,\text{m/s}$):

$$f_{\text{cut-on}} \approx 5.3\,\text{kHz}\,.$$

Hence, we must be careful not to ascribe too much credence to predictions of the vocal-tract acoustics for frequencies above 5 kHz.

2.3.6 End corrections

Although the vocal tract has smooth changes in the geometry function along most of its length, there are some locations at which the area changes abruptly, such as at the pyriform sinuses, the teeth and the lips. When the area changes abruptly, an end correction is required to account for the disparity between the acoustics of the idealised model geometry and reality. End corrections are a simple practical means of incorporating some spatial aspects of real duct acoustics into a one-dimensional model, and are required to modify the model geometry so that the predicted acoustic behaviour using plane-wave theory alone closely matches the observed response of real systems. At abrupt changes in the cross-sectional area, the transfer of the acoustic pressure from one tube section to a wider one, e.g., from 1 to 2 in Figure 2.6, responds to the additional acoustic inertance as if the narrower tube were slightly longer. Hence, the first tube is extended by an end correction ϵ that is calculated from the tube areas, S_1 and S_2, changing the point of transfer between the two sections: $x_1 = l_1 + \epsilon$; $x_2 = l_2$. The formula for computing the end-correction factors is based on calculations and empirical results from rigid-walled tubes (Davies et al. 1980; Davies 1988, Eq. 4.10, p. 104):

$$\epsilon = \kappa r_1 \left[1 - \exp\frac{2}{3}\left(1 - \sqrt{\frac{S_2}{S_1}}\right)\right], \tag{2.14}$$

where $\kappa = 0.63$, and r_1 is the hydraulic radius in tube 1. (This expression is similar to that for an open end, which is given in Appendix A.6.)

2.3.7 Side branches

Side branches occur at various points in the supralaryngeal vocal tract, for instance the nasal and pyriform sinuses. It may also be advantageous to model other geometrical discontinuities as side branches, such as the sublingual cavity. Within the framework of a plane-wave model,

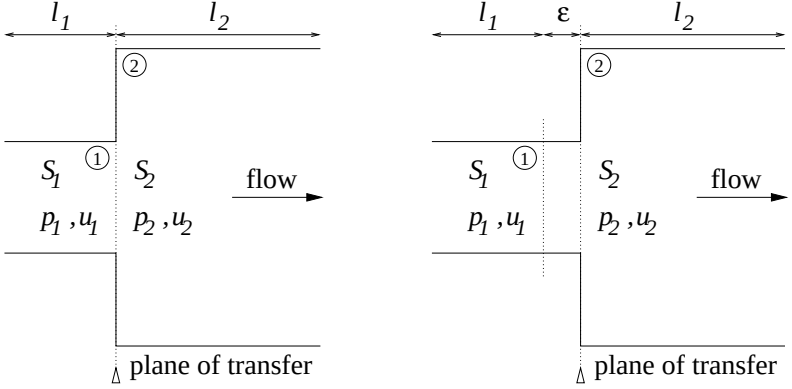


Figure 2.6: The physical geometry (left) and its representation within VOAC (right), showing the lengthening effect of the end correction on the narrower tube and the new position of the plane of transfer.

tube sections can be added facing either upstream or downstream. More sophisticated models might allow the side branch to be added at an arbitrary angle (e.g., Dang et al. 1997), but that is not considered here. In the model, parallel side branches can be treated as a special case of the end correction where the tube of section 1 extends into tube 2, and the pressure transfer takes place at the interface, as before, as illustrated in Figure 2.7. Now, with two abrupt changes, from 1 to 2 and from 3 to 2, end corrections must be calculated for tube 1 and for tube 3. To solve for the partial pressures in all three sections, an extra condition must be used beyond mass and momentum: in tube 3, the reflection coefficient at the closed end yields a relation between p_3^+ and p_3^-. When flow is included, a further condition is required to resolve the entropy changes at the plane of transfer, which is derived by conservation of energy (further details are given in Appendix A).

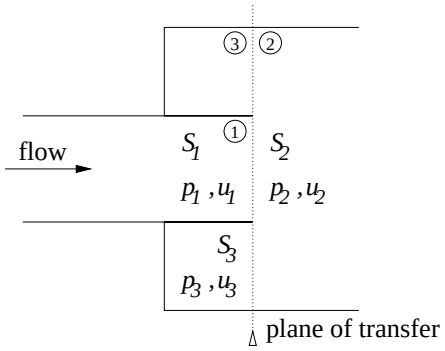


Figure 2.7: Expansion geometry with a side branch showing sections 1, 2 and 3.

2.3.8 Flow

The most significant extension of VOAC beyond traditional approaches is in the inclusion of net flow. Not only does it account for the changes to propagation time in a moving fluid, but

it allows for flow separation, jet formation and turbulent mixing, without any departure from the plane-wave paradigm. For low Mach numbers there is very little effect in terms of losses or changes to resonance frequencies, but substantial flow velocities are quite common in speech, particularly during frication or the release of a burst, often reaching values $M \geq 0.3$. The details of the derivation and implementation of the equations for flow are given in Appendices A and B respectively, but examples of the effects of flow will be presented against measurements from flow duct experiments in the process of validating VOAC later in this chapter (Section 2.5.2).

2.4 Implementation

To enable VOAC to cope flexibly with a variety of different vocal-tract geometries, it is necessary to construct the model of the flow duct from a number of geometrical primitives. In this section, we describe how these are implemented, how the program has been modified so that it can be used to predict the transfer function for non-terminal acoustic sources, and various other details of the implementation.

2.4.1 Element types

The geometry function, which is the shape information required for acoustic predictions, is the axial distribution of vocal tract area (area function) and cross-sectional shape (hydraulic radius function). It is divided into a set of discrete sub-elements, each of which is equivalent to a single section of tube. For RAMPs and CONEs, the area of one sub-element varies smoothly along its length, while for the other sub-element it stays constant. Other type of element are constructed purely from constant-area sub-elements. Using various combinations of these sub-elements provides us with considerable flexibility in defining an accurate representation of the vocal-tract geometry from the available anatomical information. The sub-elements are grouped into elements that can be one of five inherited types (see Figure 2.8): 1. ORIFICE, 2. RAMP, 3. CONE, 4. OUTLET (with side-branch option), or 5. PIPE. Each type incorporates a different function of the cross-sectional area $S(x)$ with respect to distance x, which is defined at junctions j: x_j, S_j. Nonetheless, they all require that any change in area should occur within the element, and not between element boundaries. This condition ensures that all pressure transfers take place within each element, even after end corrections have been applied.

Each ORIFICE or OUTLET element normally contains a contraction then an expansion, or vice-versa, although the second abrupt area change is optional and may be left undefined. Thus, the ORIFICE (Type 1) always accommodates an area expansion ($S_2 < S_1$). The RAMP and CONE each comprise a single gradual area change, either linear or quadratic with distance, plus an optional constant cross-section tube, identical to the PIPE element itself. The RAMP

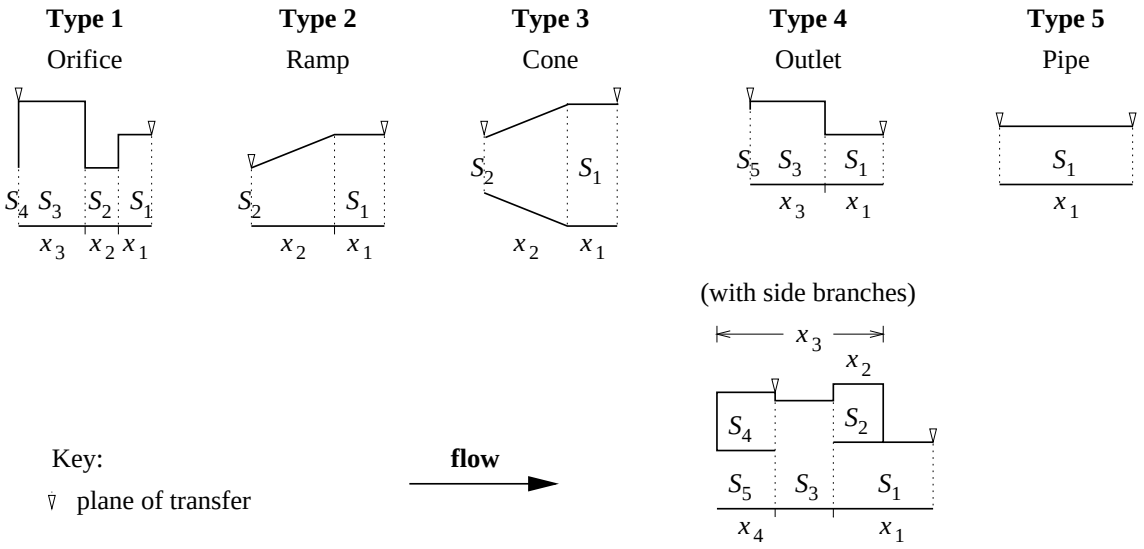


Figure 2.8: The choice of five element types in VOAC, showing Type 4 with and without side branches. The lengths along the tract are denoted by x_j and the cross-sectional areas by S_j at the planes indicated by the dotted lines.

(Type 2), which has a linear change in area, assumes cylindrical wave propagation for the transfer calculation; the CONE (Type 3), which is in fact a frustrum or truncated cone, has a quadratic area function and assumes spherical wave propagation within that section. The OUTLET (Type 4) has the option of including side branches facing either upstream or downstream, to simulate sinuses for example, in addition to an increase in area ($S_3 > S_1$). The PIPE (Type 5) maintains a constant area along its length. Each element can contain fewer sections than the illustrations, but not more. For instance, Type 1 may have only two tube sections if $x_3 = 0$, and no contraction if $S_3 = S_2$. Figure 2.9 gives Fant's /ɨ/ as an example of the area and hydraulic radius functions with respect to distance along the vocal tract (Fant 1960). It is constructed using Types 1, 3, 2, 2, 1, 3 and 1, starting at the lips.

2.4.2 Supraglottal sources

For a model that is linear in sound pressure and time-invariant with respect to its reverberation time, it can be shown that the transfer function from an intermediate source in the vocal tract to the lips, as depicted in Figure 2.10, is equal to the transfer function from the glottis to the lips, divided by that from the glottis to the source location (Henke 1966; Liljencrants 1985). Indeed, Shadle (1985) showed that the poles of the whole VTTF, i.e., of $H^V_{GL}(\omega)$, were poles of the source-lips TF $H^P_{QL}(\omega)$, and that the poles of the rear-tract TF $U_G(\omega)/p_Q(\omega)$ were system zeros of $H^P_{QL}(\omega)$, provided there were no blockages or ports along the tract (i.e., for finite series impedances and shunt admittances, see Shadle 1985, Section 3.1.2 from p. 72, Section 3.1.6 from p. 102, and Appendix B pp. 184–5). For an ideal source, the TF from the source location to the

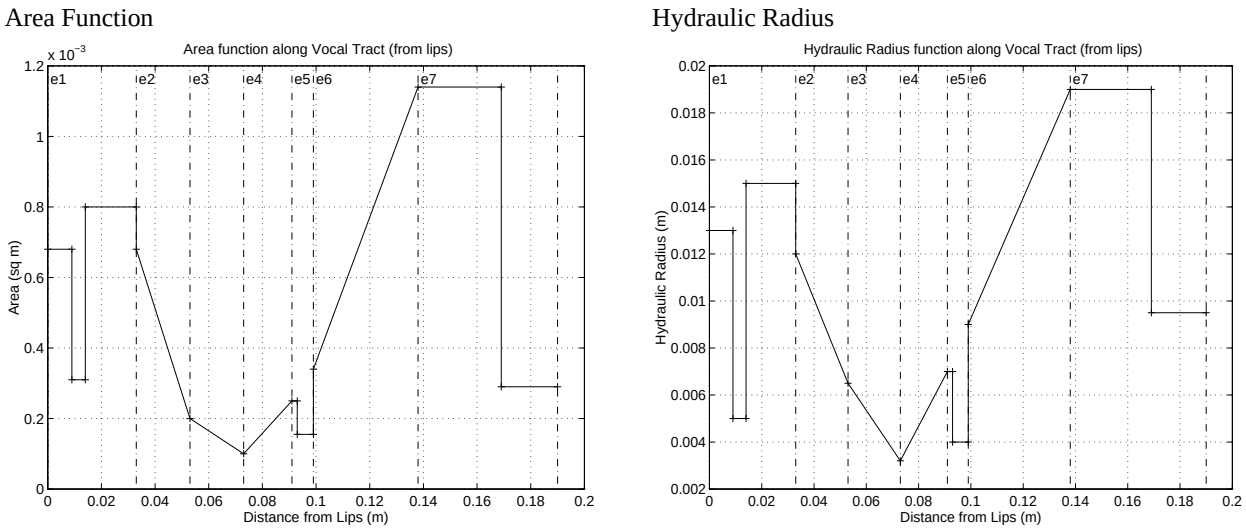


Figure 2.9: Area function and hydraulic radius for Fant's /i/. The labels (e1, e2, etc.) indicate the element number (*not* type number), and the dot-dashed lines mark the boundaries between the elements.

glottis, which is the rear-tract TF, requires a reflection coefficient at the source of $R = 1$ for a pressure source, or $R = -1$ for a volume-velocity source. In other words, the partial pressures at Q are equal, $p_Q^+ = p_Q^-$, for a pressure source ($p_Q = p_Q^+ + p_Q^-$), and opposite, $p_Q^+ = -p_Q^-$, for a volume-velocity source, $U_Q = (p_Q^+ - p_Q^-)S_Q/\rho_0 c_0$, where S_Q is the cross-sectional area at the source location.

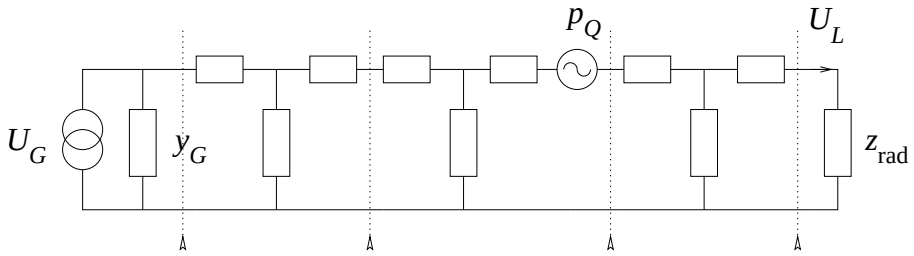


Figure 2.10: Transmission line representation of a supraglottal source. Each dotted line rising from an arrowhead denotes a junction between tube sections.

Writing the TF between two points A and B as $H_{AB}(\omega)$ for volume velocity $U(\omega)$ and pressure $p(\omega)$, and using G, L and Q to denote the location of the glottis, the lips and the source respectively, the overall VTTF from a pressure source to the volume velocity at the lips is defined as:

$$H_{QL}^P(\omega) \quad = \quad \left(\frac{U_L(\omega)}{U_G(\omega)}\right)_{R=\frac{z_{\rm rad}-1}{z_{\rm rad}+1}} \left(\frac{U_G(\omega)}{p_Q(\omega)}\right)_{R=1} \tag{2.15}$$

where the volume-velocity VTTF is $H_{GL}^V = U_L/U_G$, and the pressure VTTF is $H_{QG}^P = U_G/p_Q$. This relation is proven in Appendix A.7 for a pressure source part-way along a simple tube that is closed at one end, as shown in Figure 2.11.

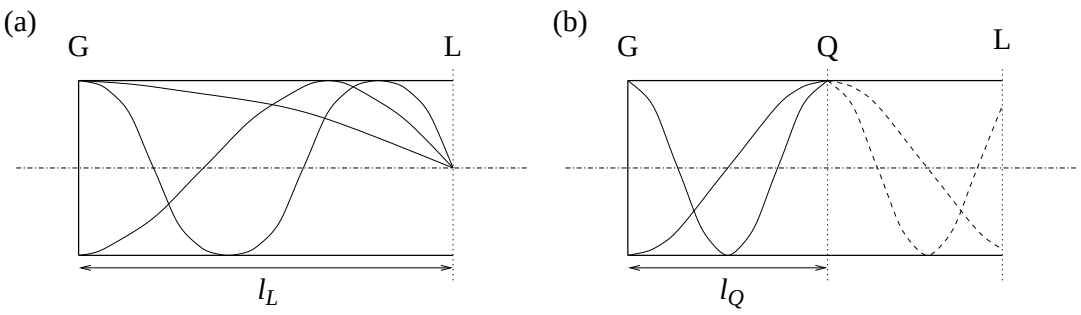


Figure 2.11: The pressure modes for a simple tube that is closed at the left-hand end, for (a) the whole tube and (b) the rear-tract excited by a pressure source at Q.

By means of illustration, let us consider the poles and zeros of the above example (Fig. 2.11). Hence, ignoring the effects of sound radiation, the frequencies of the standing plane-wave modes Fi can be calculated approximately by the formula

$$\mathrm{F}i = (2i-1)\frac{c_0}{4l_L}\,, \tag{2.16}$$

where i is any positive integer excluding zero, c_0 is the speed of sound and l_L is the length of the tube. These modes appear as resonances or poles in the VTTF, whereas for a source at point Q within the vocal tract, the resonances of the rear-tract are the system anti-resonances or zeros. For a pressure source at a distance l_Q from the closed end, the boundary condition at the source is equivalent to another closed end, since the reflection coefficient is $R = 1$. Hence, the zeros occur at frequencies Zi of the even modes:

$$\mathrm{Z}i = i\frac{c_0}{2l_Q}\,, \tag{2.17}$$

but in this case, i can take a value of zero ($i \in \{0, 1, 2, \ldots\}$), resulting in a net low-frequency anti-resonance. A schematic depiction of the frequency response for a pressure source at $l_Q \approx 0.6l_L$ is given in Figure 2.12, which includes some losses from sound absorption and radiation.

As already mentioned in Section 2.2.5, an intermediate source has been added to the list of source types that can be modelled by VOAC, which was achieved by a two-stage calculation of the terms in Eq. 2.15: $H^V_{GL}(\omega)$ and $H^P_{GQ}(\omega)$. First, the volume-velocity TF of the complete geometry function was computed in the normal way and, second, VOAC was re-run to compute the TF from volume velocity at the glottis to pressure at the source, using just the rear-tract. However, the acoustic velocity is unaffected by an ideal pressure source, which implies a reflection coefficient $R = p^-/p^+ = 1$. Therefore, for the second stage, in place of the usual radiation impedance, an infinite impedance (open circuit) was presented at the source location to the rear-tract. The two TFs were then combined to yield $H^P_{QL}(\omega)$, the TF of the volume velocity at the lips from an intermediate, or supraglottal, pressure source.

If we assume that placing a source at one location does not alter the transfer function from another, cases of mixed sources can be modelled by simply summing the vocal-tract responses

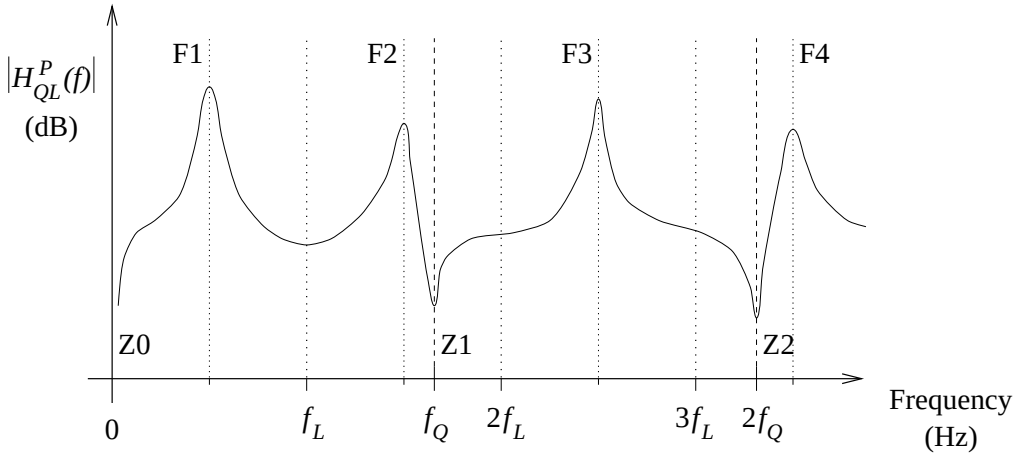


Figure 2.12: Sketch of the magnitude of the transfer function H^P_{QL} versus frequency, from a pressure source p_Q that is a distance l_Q from the closed end of a simple tube of length l_L, to the volume velocity at its open end. The frequencies marked on the horizontal axis are at multiples of $f_L = c_0/2l_L$ and $f_Q = c_0/2l_Q$.

for each source. Distributed sources can be treated in much the same way — by superposing a cluster of several weaker sources acting at different places. Any interaction between sources, however, must be modelled separately, such as in a voiced fricative which has a voiced source at the glottis and a modulated frication source near the supraglottal constriction. Despite this, the same procedure remains valid for interacting sources, provided the sources do not affect the transfer functions, that is, that there is no source-filter interaction. In practice, there is always some coupling between the source and its TF, which is a major drawback of the source-filter representation. Studies of the formant resonances during voicing have shown that the variable coupling of the subglottal airways modulate their centre frequencies and bandwidths (Rothenberg 1981; 1983; Titze and Story 1997). However, these effects are relatively small ($< 100\,\mathrm{Hz}$) and can be dealt with on a piecewise basis, as necessary (Yegnanarayana and Veldhuis 1998). Using VOAC's representation of the vocal tract, both pressure and volume-velocity acoustic sources can be associated with an element (or section of an element), and in doing so provide a means for generating a distributed source structure.

2.4.3 Losses

There are many sources of energy dissipation in human speech production: at source, from flow, from wall absorption and vibration, and from radiation (see Section 2.3.4). VOAC enables us to take care of each of these losses. In the case of wall vibration, we used the recommended values from Data Sheet 1 (Davies 1991, cf. McGowan 1992), listed in Table 2.1.

Flow losses are dealt with by consideration of the energy terms at the transfer following an expansion, in terms of entropy increases from turbulent mixing. The derivation of these terms

m	mass	$21\,\mathrm{kg/m^2}$
λ	damping	$10^4\,\mathrm{kg/m^2 s}$
ω_n	natural frequency	$220\,\mathrm{rad/s}$

Table 2.1: Parameter values for the mass, damping and natural frequency properties of the vocal-tract wall, as used in VOAC.

is presented in Appendix A.

2.4.4 Vocal-tract transfer functions

The transfer functions of the vocal-tract's acoustic response are calculated at frequencies $f = \omega/2\pi$ from the pressure components at either end of the given geometry:

$$\begin{aligned} H^V_{GL}(f) &= \frac{U_L(f)}{U_G(f)} \\ &= \frac{u_L S_L}{u_G S_G}\,\frac{\rho_0 c_0}{\rho_0 c_0} \\ &= \frac{S_L}{S_G}\,\frac{\left(p^+_L - p^-_L\right)}{\left(p^+_G - p^-_G\right)}\,; \end{aligned} \tag{2.18}$$

$$\begin{aligned} H^P_{QG}(f) &= \frac{U_G(f)}{p_Q(f)} \\ &= \frac{u_G S_G}{p_Q} \\ &= \frac{S_G}{\rho_0 c_0}\,\frac{\left(p^+_G - p^-_G\right)}{\left(p^+_Q + p^-_Q\right)}\,, \end{aligned} \tag{2.19}$$

where S_L and S_G are the cross-sectional areas at the lips and the glottis, respectively; ρ_0 and c_0 are the time-averaged density and speed of sound; the superscripts $^+$ and $^-$ refer to the positive- and negative-travelling wave components, respectively. The output is a vector of complex amplitudes corresponding to the frequencies at which the response was specified.

2.5 Comparison with experiment

Preliminary tests, performed during the commissioning of the translated VOAC program, and basic evaluations using acoustic theory of simple tube configurations are detailed in Appendix B.1, but this section describes some comparisons made against experimental data. VOAC was used to predict the radiated sound spectra for some physical models, which were compared with their measured sound spectra.

2.5.1 Physical models

Apparatus

A series of flow experiments was conducted by Shadle (1985) using physical duct models, with each specimen representing a highly idealised fricative configuration. The test rig, described in detail in Section 2.1 of Shadle (1985), comprised a source of laminar air flow, the test specimen and a baffle. A tank of compressed air supplied air to the specimen via a flow regulator and a silencer at a pressure that was monitored by a manometer. The radiated sound was measured by a microphone (B & K 4133), amplified and fed into a spectrum analyser, whose output was logged digitally. Shadle's objective was to obtain accurate measurements of turbulent flow noise with each physical model specimen, and a control measurement of the apparatus with no specimen, which consisted of just a baffle at the jet exit with the obstacle positioned downstream. By assuming a semi-infinite space, the control measurements of the source characteristics (with a baffle but no specimen) were used to derive source functions for a given obstacle in the path of the jet over a range of flow rates. Measurements were made of the sound radiated from each specimen with the given obstacle inside. The TF was predicted from knowledge of the well-defined source location and dimensions of the specimen. It was combined with a regression fitted to the measured source function and the radiation characteristic (Eq. 2.4) to give the predicted sound, which was then compared with the experimental results. We have used the results from these experiments to compare VOAC's predictions against measured data.

The specimens used in these tests were physical models with geometries that were deliberately simple for purposes of acoustic modelling. They comprised a tube of fixed length, an inserted constriction and obstacle. Different specimens were created by selecting the constriction and obstacle from an arsenal of various shapes and sizes, and by adjusting their position along the tube. For the comparisons in the present study, we used the results from two geometries: specimen 1 and specimen 2.

Features of the TFs

As shown in Figure 2.13, the distance from the obstacle to the constriction l_o was kept constant, as was the entire configuration except the distance from the glottis to the constriction l_b, which was either 12.8 cm or 4.0 cm. In both cases, therefore, the TF of the whole tract and that of the rear-tract (defined as the part upstream of the obstacle) change. If we assume that the part upstream of the constriction is only weakly coupled to the part downstream, which contains the obstacle, we would expect the resonances to correspond to the (odd) modes of the downstream part, and the anti-resonances to those (even modes) of the constriction-obstacle section. Therefore, the system poles of the two specimens would differ (Eq. 2.16), but not the

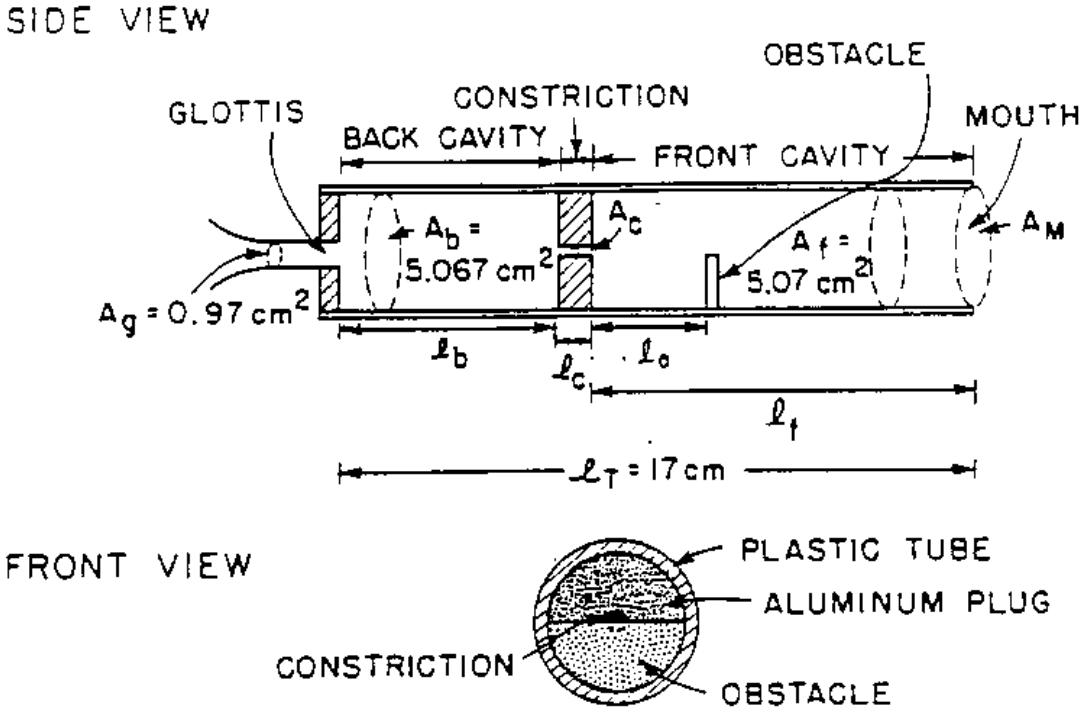


Figure 2.13: Diagram of physical flow-duct model used to form specimen 1 ($l_f = 3.2\,\text{cm}$) and specimen 2 ($l_f = 12.0\,\text{cm}$) from Shadle (1985, p. 33). For both specimens, $l_c = 1.0\,\text{cm}$ and $l_o = 3.0\,\text{cm}$.

Specimen	l_L (cm)	Fi (kHz)	Zi (kHz)
1	3.2	2.7, 8.0, 13.4, ...	0.0, 5.7, 11.4, ...
2	12.0	0.71, 2.1, 3.6, 5.0, 6.4, ...	0.0, 5.7, 11.4, ...

Table 2.2: Resonance and anti-resonance frequencies, Fi and Zi, estimated by Eqs. 2.16 and 2.17 respectively, for the physical models ($c_0 = 343\,\text{m/s}$, $l_Q = 3.0\,\text{cm}$ and $l_L = l_f$).

zeros (Eq. 2.17) in this approximation, as indicated in Table 2.2. The effect of the weakly-coupled part of the tract, upstream of the constriction, is to produce many other zeros and poles that are nearly equal, and almost completely cancel each other. They appear as small kinks in the overall frequency response of the TF. The frequency values given in Table 2.2 are those of the free zeros and uncancelled poles.

Geometry functions

The geometry function of specimen 1 is plotted against the length along the tract in Figure 2.14 with a 2 cm-long inlet at the glottis. It shows both the area function and the corresponding hydraulic radius function, which differs in shape only at the semi-circular obstacle. The turbulence noise is assumed to be generated by a pressure source on the upstream edge of the obstacle, i.e., where the jet would impinge upon it.

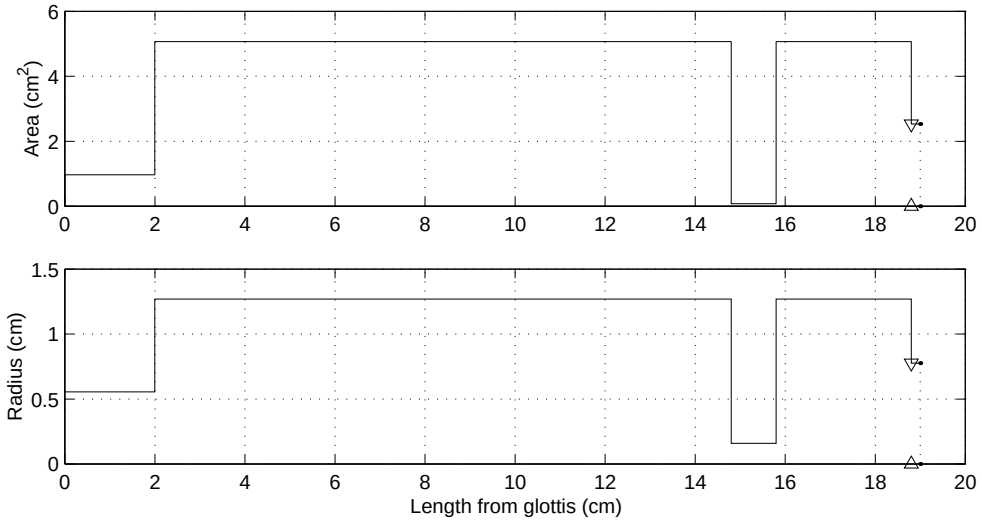


Figure 2.14: Geometry function of specimen 1, which consists of (top) the area function and (bottom) the hydraulic radius function. The radiation surface is shown as a dotted line at the right-hand end, and the source location used for the VTTF calculation is indicated by the triangles.

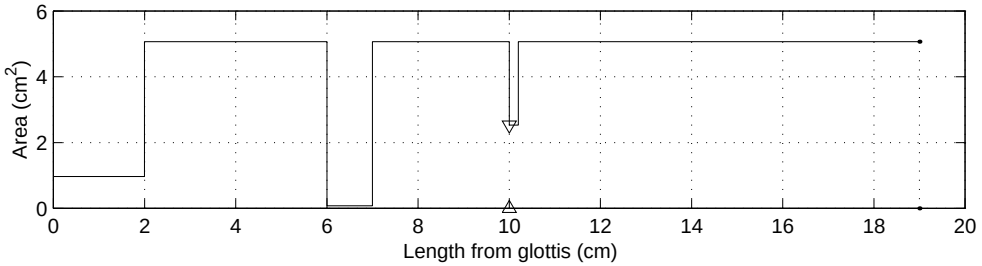


Figure 2.15: Area function of specimen 2, showing the radiation surface (dotted) and the source location (triangles).

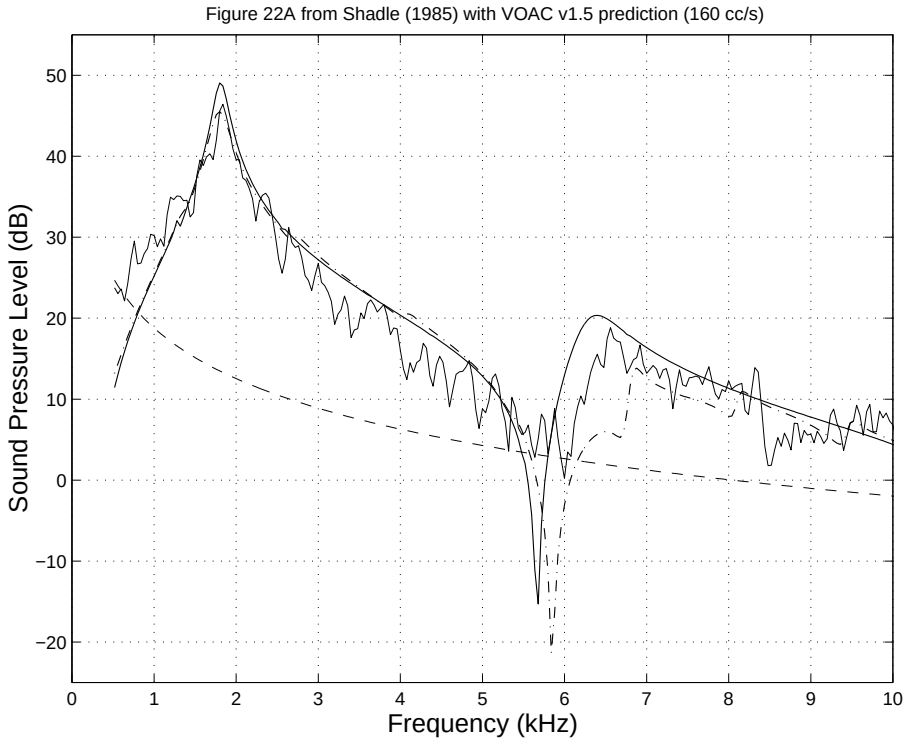


Figure 2.16: Specimen 1: measured (thin solid) and predicted sound spectra in the far field at a flow rate of 160 cm^3/s, using the CEA (thick dash-dot) and VOAC (thick solid). The thin dashed curve is the noise floor.

Figure 2.15 depicts the area function for specimen 2, for which the constriction has been moved towards the glottis. Note that the distance between the constriction and the obstruction downstream of it is identical for the two specimens.

2.5.2 Frequency response functions (FRFs)

Specimen 1

The values of sound speed $c_0 = 344.8\,\mathrm{m/s}$ and density $\rho_0 = 1.18\,\mathrm{kg/m}^3$ were set to those used by Shadle (1985) in the earlier study, for which consistent values of temperature $T = 293\,\mathrm{K}$ and the ratio of specific heats $\gamma = 1.40$ were derived. The measured sound spectrum, which is the thin solid line in Figure 2.16, is drawn above a regression of the measured noise floor (thin dashed curve). A prediction using the classic electrical analogue (CEA) was made in that study, and is drawn as the thick dash-dot line; the response that was predicted by VOAC is superimposed as a thick solid line. A summary of the estimated formant frequencies and their bandwidths is given in Table 2.3.

The VOAC predictions are within 7 dB of measurements for the lower part of the spectrum ($f < 5\,\mathrm{kHz}$), which is 10 % of the dynamic range of the predicted response, and approximately three times the deviation of the noise measurements. The first resonance F1 $\approx 1.8\,\mathrm{kHz}$ is

	(Hz)	F1	F2
Measured	Fi	1830	6550
	BW	130	200
CEA	Fi	1790	6870
	BW	270	560
VOAC	Fi	1810	6380
	BW	180	670

Table 2.3: Centre frequency (Fi) and bandwidth (BW) of the formant resonances measured, and predicted by CEA and by VOAC, for specimen 1.

well-matched in overall amplitude, frequency and bandwidth, as is the anti-resonance Z2 $\approx$ 5.7 kHz, for the part above the noise floor. There are discrepancies above this frequency, however, which can be attributed to poorer estimation of the higher modes from multiplicative errors, the influence of cross modes or other modelling inaccuracies. Even so, the predicted spectrum stays within the same error bound as the CEA prediction, which has anomalies of approximately 10 dB between 6 kHz and 7 kHz. The small blip in the VTTF predicted by VOAC at 1.3 kHz is evidence of a closely matched pole-zero pair, produced by the cavity upstream of the constriction.

Specimen 2

The results for specimen 2 are given in Figure 2.17, which also show good general agreement between the predicted TF and measurements (i.e., within $\pm$6 dB). There is some misalignment of the centre frequencies for F3 and F6, but the resonance frequencies are otherwise accurate, as seen in Table 2.4. As before, the anti-resonance, which is marginally lower at Z2 = 5.6 kHz, provides a reasonably faithful fit to the measured sound spectrum as far as the noise floor allows. The CEA also captures the main features of the spectrum, except in the 6–7 kHz region. The damping at the lower formants appears to be too low in the VOAC predictions by comparison with the measurements; however, from about F3 upwards the resonance bandwidths appear to be accurate. The story is similar for the CEA predictions, although the bandwidths predicted by CEA are generally higher.

2.5.3 Discussion

The quality of the match with the experimental data is similar for VOAC and the CEA, although it could be argued that VOAC offers a small improvement over the CEA. Significantly, VOAC has the advantage of automatically adjusting resonance frequencies and increasing losses as the flow rate was increased, and does not require ad hoc adjustment of its parameters to

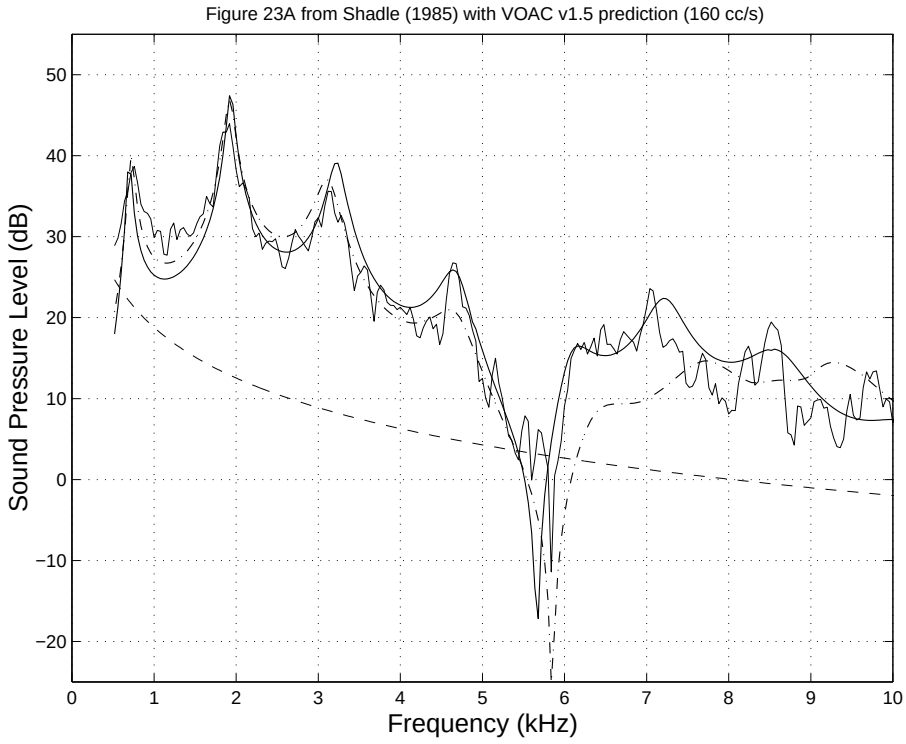


Figure 2.17: Specimen 2: measured (thin solid) and predicted sound spectra in the far field at a flow rate of $160\,\mathrm{cm}^3/\mathrm{s}$, using the CEA (thick dash-dot) and VOAC (thick solid). The thin dashed curve is the noise floor.

	(Hz)	F1	F2	F3	F4	F5	F6	F7
Measured	Fi	760	1920	3130	4650	6150	7050	8500
	BW	160	180	130	160	–	130	200
CEA	Fi	740	1920	3130	4610	6530	7740	9260
	BW	70	130	270	–	–	1070	–
VOAC	Fi	720	1920	3220	4630	6170	7200	8550
	BW	90	90	180	380	–	470	–

Table 2.4: Centre frequency (Fi) and bandwidth (BW) of the formant resonances measured, and predicted by CEA and by VOAC, for specimen 2.

fit these particular results. Note that there may have been minor errors in the positioning of the constriction and obstacle along the tract, for example, the difference between the value of Z2 predicted by VOAC (5.6 kHz) and the frequency of the measured spectral minimum (5.8 kHz) could be caused by an error of 0.1 mm. Also note that Fi are slightly over-estimated in Table 2.2, because of the absence of any radiation term, and are therefore higher than the measured formants. The end correction that would be equivalent to the radiation impedance is bigger for specimen 1 because its radiating area is smaller than for specimen 2, but since its front cavity is much smaller, the end correction may have a greater effect. The first anti-resonance at Z1, which has only a minor end correction factor, is reasonably accurate.

The narrow bandwidth of the lower formants is a result of insufficient losses in the model. The assumption of a piston in a baffle tends not to hold true for $f < 1$ kHz, which may cause the net losses to have been under-estimated for the low-frequency range. Moreover, the specimens are not perfectly rigid in practice, and we might expect wall vibration to have a more significant effect at low frequencies, for which the walls appear more flexible.

2.6 Summary

In this chapter, a frequency-domain, flow-duct acoustics program that we have revised and extended, VOAC, has been described and illustrated. VOAC uses the geometry of the vocal tract to predict the impulse response in the far field from a source anywhere within it. We have tested its output against experimental flow-noise data with the conclusion that the predictions compared well with measurements.

Many of the standard assumptions made in models of the vocal-tract acoustics were relaxed in VOAC's earlier formulation (Davies et al. 1993): net flow and changes in entropy from flow separation were allowed; both abrupt and gradual area changes were modelled using spherical, cylindrical or planar waves; the effects of cross modes were provided for by end-correction factors and sinuses could be added as side branches; the losses from wall vibration, viscosity and heat conduction were incorporated. The development of the current version has entailed translation to Matlab, the building of various input and output utilities and implementation of an intermediate source option. The utilities perform tasks such as reading geometrical data and plotting VTTFs. In the following chapter, VOAC is enlisted to display its potential for articulatory synthesis, specifically by performing speech synthesis using experimental measurements of the geometry function from magnetic resonance images.

Chapter 3

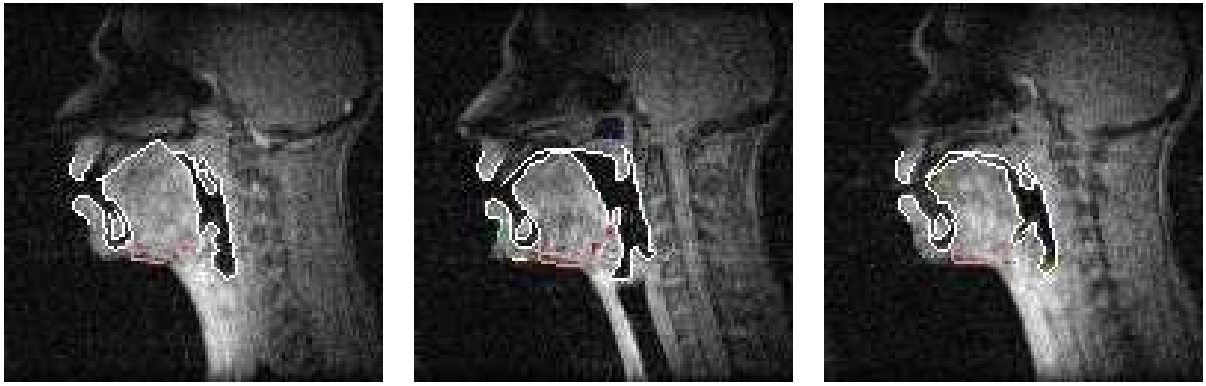
From images to sounds

3.1 Introduction

Measurements of the precise geometry of a subject's vocal-tract configuration are not usually easy to obtain, since many techniques are hazardous (e.g., X-ray) or interfere with the subject's ability to speak (e.g., EPG, electromyography or a velar trace). In recent years, magnetic resonance imaging (MRI) has become much more accessible to those conducting speech research and, since it has no known side-effects, MRI studies have proliferated. By capturing more than one image slice, MRI acquires three-dimensional data, which are needed to quantify the full vocal tract geometry. Numerous studies have used MRI on the vocal tract to derive the vocal-tract area function while the subject sustained particular phonemes (Baer et al. 1991; Beautemps et al. 1995; Narayanan et al. 1995; Alwan et al. 1997; Story and Titze 1998b). In the present study, we have used only three-slice sagittal data that were available for subject PJ, because they were for the same subject used for the speech recordings analysed in later chapters. An important aspect of these data is the high frame rate, obtained by averaging successive repetitions of the word in question and aligning the frames according to the acoustic signal. We refer to this technique as dynamic MRI, or dMRI. In this chapter, the process of interpreting the dMRI frames to produce area functions and of predicting the sounds that these vocal-tract configurations produce are described, using data gathered by a related project (Mohammad 1997). Results are shown for two vowels [ɑ] and [i], a fricative [s], and a plosive [p^h], taken from the nonsense word /pɑsi/, and then compared to analysis of corresponding speech recordings from the same subject.

3.2 The dMRI data

This section describes how the raw image data files were gathered in the hospital, and how the outline of the vocal tract was marked on each of the images.



Left Mid-sagittal Right

Figure 3.1: Sagittal dMRI slices, left, middle and right, for the vowel [i] in [p^hɑsi] by PJ (frame 31). The segmented outlines (white) are overlaid, which include the lower mandible but not the teeth.

3.2.1 Acquisition

The raw image data files were acquired as part of a collaborative project to explore improvements in the time resolution and the combination of several slices to give volumetric data, which formed the core of Mohammad's PhD thesis (Mohammad 1999). A 0.5 T SIGNA GE scanner was set to scan using fast RF-spoiled gradient echo, and was programmed to save the interleaved raw data to file, using a spatial resolution of 1 pixel = 1.875 mm×1.875 mm. Data files were made from a sequence of 24 scans captured during hundreds of repetitions of the nonsense word /pɑsi/ spoken by an adult male (PJ), who is a native speaker of British English RP. By synchronising the image data to acoustic cues from simultaneous recordings, images were reconstructed, which effectively reduced the time resolution to 16 ms. Three 5 mm-thick, sagittal slices were taken, spaced 11 mm apart and centred on the mid-sagittal plane, which provided a source of three-dimensional dMRI data: left, middle and right, as shown in Figure 3.1 for the mid-phoneme frame of the vowel [i].

3.2.2 Segmentation

The sagittal images generated from the raw data were manually segmented with the outlines of the pertinent anatomical features: the upper lip, the hard palate, the soft palate and velum, the back wall of the pharynx (including the pyriform sinuses), the vocal folds, the epiglottis, the tongue body, the lower mandible (i.e., jaw bone) and lower lip. These outlines are shown superimposed on the images in Figure 3.1.

The upper and lower teeth, although not always clear from the images, were also superimposed on the images by a combination of careful manipulation of the images (e.g., by histogram equalisation), reference to other parts (e.g., mandible, lips and tongue) and subjective judge-

ment. The nasal cavity was ignored. The complete vocal-tract outlines were exported as a chain of connected pixel-centres on a fixed reference grid for the next stage of interpretation. The locations of the selected pixels were deemed to be accurate with 99 % confidence, and so the standard deviation of the error in the outlines was taken to be nominally $\pm\frac{1}{3}$ pixel (Mohammad 1999). These chains of pixel coordinates were linked sequentially to provide a single contour describing the outline of the vocal tract. The vocal-tract outline derived from the middle slice for the mid-phoneme frame of the vowel [i] is shown in Figure 3.2.

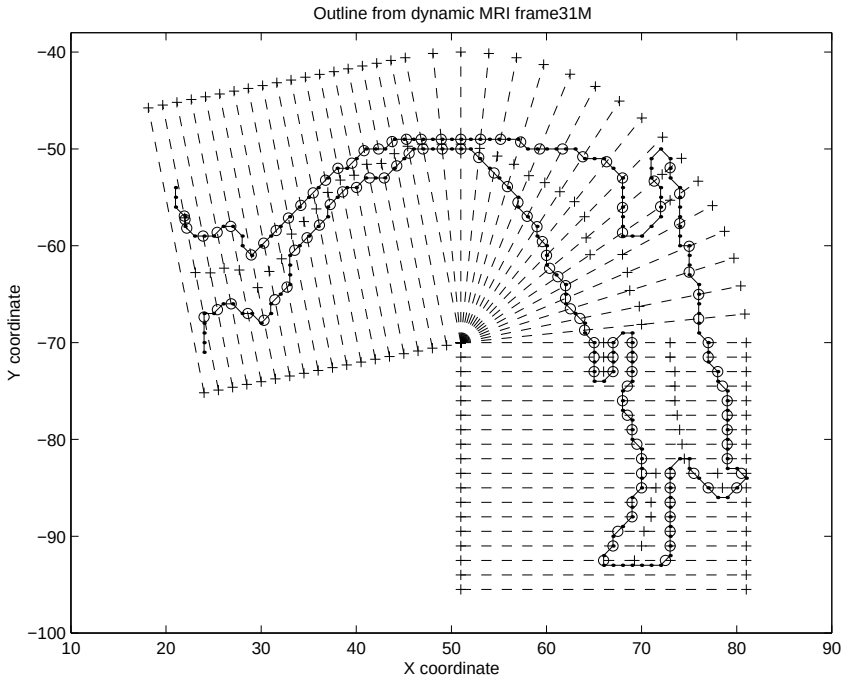


Figure 3.2: Double-density grid (dashed lines) overlaying the outline from the mid-sagittal slice (dots joined by a solid line) for [i] spoken by PJ (frame 31). The intercepts are shown by circles, and crosses mark the mid-point of each vocal tract section, as well as the ends of grid lines.

The upper and lower boundaries were sometimes coincident in the outlines, when, for example the tongue was touching the roof of the mouth, as seen in the left-hand slice in Fig. 3.1 (left). In these cases, the position of the outline was somewhat arbitrary, although the hard palate's profile was reasonably stable over the many frames. Nevertheless, the outlines clearly represent the movement of the principal articulators and enabled us to derive a description of the vocal-tract geometry, in terms of the cross-sectional area and the hydraulic radius (as stated in Chapter 2).

3.3 Distance functions

Conversion from the three outlines into a single geometry function for each frame was performed via the vocal-tract cross-sectional distances. The method of converting each vocal-tract outline

to a profile of distances, or distance function, is presented in this section. The last part of the process is detailed in the following section, where distance functions are converted into geometry functions that comprise profiles of area and of hydraulic radius along the tract.

3.3.1 Overlaying a grid

Distance functions were generated by taking a series of measurements, as defined by a grid laid over each outline. Initially a series of pixel-quantised lines was overlaid on the processed image to identify the coordinates of the intercepts, but the coarse resolution of the pixels made this approach problematic for slanted lines. So, an alternative was adopted which connected the pixel centres to make a continuous contour along the outline. Thence, a coarse grid was drawn consisting of a series of parallel horizontal lines in the lower vocal-tract region, radial lines (centred at the tongue centroid) around the top of the pharynx and the back of the oral cavity, and slanted parallel lines running along to the lips. The parallel lines were originally set three pixels apart in the vertical and horizontal directions, respectively. The radial lines were $\pi/16$ radians apart (dividing a right angle into eight segments), to give a spacing in the vocal tract comparable to that of the parallel lines, and an additional line was included past the vertical, which resulted in an angle $\theta = 9\pi/16$, i.e., greater than 90°. The downward slope of the slanted lines provided a better cut for the cross-sections of the anterior oral cavity, which tended to decline before levelling out towards the lips.

To assess the effect of discretisation of the vocal-tract distance functions, finer grids were drawn by multiplying the number of grid lines by 2, 3 and 6, resulting in interline separations of $1\frac{1}{2}$ pixels ($\pi/32$ rad), 1 pixel ($\pi/48$ rad) and $\frac{1}{2}$ pixel ($\pi/64$ rad), respectively. Although a small amount of information was lost in the discretisation process, the double-density grid was deemed to be sufficient, after consideration of the size of errors on the outlines and informal evaluation of the acoustic consequences of the quantisation errors. This grid comprised a set of horizontal lines $1\frac{1}{2}$ pixels apart (2.8 mm), 18 radial lines separated by $\pi/32$ rad and another set of parallel lines declining at $\pi/16$ rad, as shown in Figure 3.2. The full set of outlines for the left, middle and right slices from each of the four phones [p, ɑ, s, i] are shown in Figure C.3 in Appendix C.

3.3.2 Finding the intercepts

The intercepts of the grid with the outlines were found by identifying the pair of outline coordinates that crossed the gridline, and then linearly interpolating to the point of intersection. The crossing coordinates, (x_1, y_1) and (x_2, y_2), were the ones where the angle subtended to one end of the gridline changed sign, relative to the angle of the line θ (being careful to avoid phase-wrapping artefacts). If we take the end of the gridline (lower left) as the origin, as illustrated

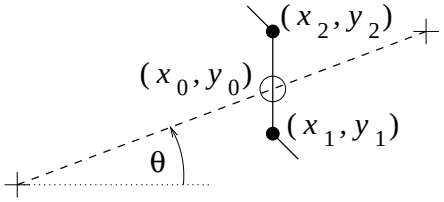


Figure 3.3: Sketch of a part of the vocal-tract outline (solid), illustrating the identification of a point of interception (Ø) at (x_0, y_0), by interpolation between the outline coordinates (•), (x_1, y_1) and (x_2, y_2), that straddle the grid line (dashed, +).

in Figure 3.3, the coordinates of the intercept (x_0, y_0) can be written:

$$x_0 = x_1 + kx_2\,, \tag{3.1}$$

$$y_0 = y_1 + ky_2\,, \tag{3.2}$$

$$\text{where } k = \frac{(x_1 \tan\theta - y_1)}{(y_2 - x_2 \tan\theta)}\,.$$

The cross-sectional distance, therefore, is simply the Euclidean distance between intercepts of the same gridline: $\sqrt{{\Delta x_0}^2 + {\Delta y_0}^2}$. Then, by combining knowledge of the direction of crossing with the position around the outline, side branches and the main cavity can easily be disambiguated.

An example of the process is illustrated in Figure 3.2 with a mid-sagittal outline for the vowel [i]. The intercepts have been circled, and a cross (+) has been placed at the centre of each sectional distance identified. When there is an odd number of intercepts, the correct closed pair has been chosen and the extraneous point discarded. Many elaborate methods have been devised to determine the vocal tract centreline and the piecewise lengths of vocal-tract elements, but ours was relatively straightforward. The length along the vocal tract l_i was defined as the perpendicular distance between parallel grid lines and as the length of the arc for radial lines, using the mid-point to define the effective radius r_i:

$$l_i = l_{i-1} + \frac{\pi}{32} r_i \qquad \text{for } 18 \le i < 36\,. \tag{3.3}$$

Side branches were also identified and their details stored with the place and sense of attachment to the main tract. A distance function computed in this way is given in Figure 3.4, showing the main cavity (above the x-axis) and side branches (below).

This mid-sagittal, mid-vowel [i] frame follows a gradually varying profile over most of its length, but has discontinuities at each of the side branches and, to a lesser extent, at the teeth (c. 15 cm from glottis). Of the four side branches, the smallest additional contribution, which is at the lips is the result of pixel quantisation and may be discarded. Starting from the glottis, the other three correspond to the pyriform sinuses, the epiglottis and the velum, as can be seen

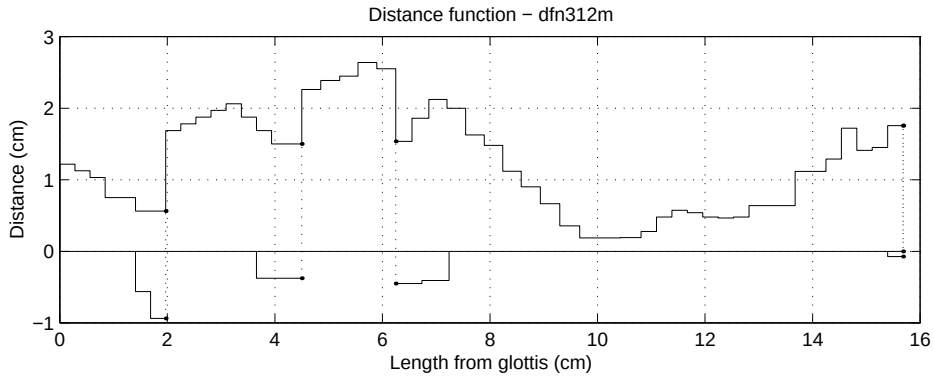


Figure 3.4: Distance function for the mid-sagittal slice of [i] spoken by PJ (frame 31). The main tract is drawn above the x-axis, whereas any side branches are shown beneath.

in Figure 3.2. Although increasing the resolution of the overlaid grid reduces the size of the steps in distance along most of the tract, abrupt discontinuities remain.

3.4 Conversion into geometry functions

Distance functions, whether single slice or multi-slice, naturally provide an incomplete description of the vocal-tract geometry, but by making certain assumptions, we can at least obtain representative geometry functions that can be supplied as one-dimensional input to VOAC. Published sources of geometric data tend to be in the form of area functions (Fant 1960; Baer et al. 1991; Narayanan 1995; Story and Titze 1998a; Story and Titze 1998b), but sometimes include the mid-sagittal distance (Beautemps et al. 1995). From these various forms of data we need to determine some rules for generating both the area and hydraulic radius profiles that are required for geometry functions for VOAC. We will begin by describing the simplest and then introduce gradually more sophisticated methods, which are designed to give more realistic results (specific to the vocal tract).

If the given data source contains only the mid-sagittal distances D or the areas S, then assuming a circular cross-section enables the hydraulic radius r (and the area) to be calculated trivially:

$$r = \sqrt{\frac{S}{\pi}}, \tag{3.4}$$

where $S = \pi D^2/4$. If both the area and the mid-sagittal distance are available, we can combine them to estimate the hydraulic radius, using an elliptical approximation. The area of an ellipse is equal to πab, where a and b are its axes, and its perimeter is approximately $2\pi\sqrt{\left(a^2+b^2\right)/2}$. Therefore, if we take half the mid-sagittal distance as the length of one axis, we have,

$$r = \frac{2S}{2\pi}\sqrt{\frac{2}{\left(\frac{2S}{\pi D}\right)^2 + \left(\frac{D}{2}\right)^2}},$$

$$= \frac{2\sqrt{2}\,S\,D}{\sqrt{16S^2 + \pi^2 D^4}}\,. \tag{3.5}$$

Now, in calculating the area functions, Fant (1960) used estimates of the cross-sectional shape at a number of stages along the vocal tract, based on inferences from anatomical data, to augment the information extracted from the sagittal X-ray images. In contrast to this purely empirical approach, Beautemps et al. (1995) devised a numerical scheme for capturing the characteristics of the differing profiles using a non-linear combination of coefficients which varied smoothly with distance along the tract x (Badin et al. 1995; Beautemps et al. 1995). The coefficients, $\alpha_{\mathrm{inf}}(x)$ and $\alpha_{\mathrm{sup}}(x)$, were optimised for a given subject by minimising the difference between measured formants, extracted from the power spectrum of speech recordings, and those predicted from the derived area function. They were composed of a spatial Fourier series for the vocal tract, and thus the smoothness of α_{inf} and α_{sup} was controlled by the number of terms, which was restricted to four (i.e., the mean value plus three sinusoids). Finally, the area was calculated according to the Heinz and Stevens model (1965):

$$S = \alpha\,(D, x)\,D^{\beta} \tag{3.6}$$

where α was a saturating interpolation of $\alpha_{\mathrm{inf}}(x)$ and $\alpha_{\mathrm{sup}}(x)$, and the exponent constant β was set to 1.5.

Figure 3.5 compares two area functions: one derived using the Beautemps method; the other assuming a circular cross-section. The hydraulic radius was determined by applying the elliptical assumption described earlier to the calculated area function, using the measured distance function.

3.4.1 Multiple slices

When more than one image slice is available, the additional information can be used to improve the quality of area estimates. In the present study, three slices were used and their distances combined in a weighted sum to yield the cross-sectional area S. Their combination can be conceptualised in two ways, as depicted in Figure 3.6: either (i) as blocks whose height depends on the distance and width on the weighting, or (ii) as a polygon connecting the ends of bars whose height again depends on distance and their spacing on the weighting. In neither case do we use the concept further; the perimeter used to calculate the hydraulic radius was derived, as before, from the area and mid-sagittal distance under an elliptical assumption. The choice of weights is governed by the width of the image slices, the inter-slice spacing and the confidence with which each outline was established, with reference to the human anatomy. Since we generally have greater confidence in the fidelity of the mid-sagittal slice, we adopted a biased set of weights, whose mean was equal to the interval between the slice centres: 9 mm, 15 mm and 9 mm for left, middle and right, respectively (shown using 10 mm, 13 mm and 10 mm in

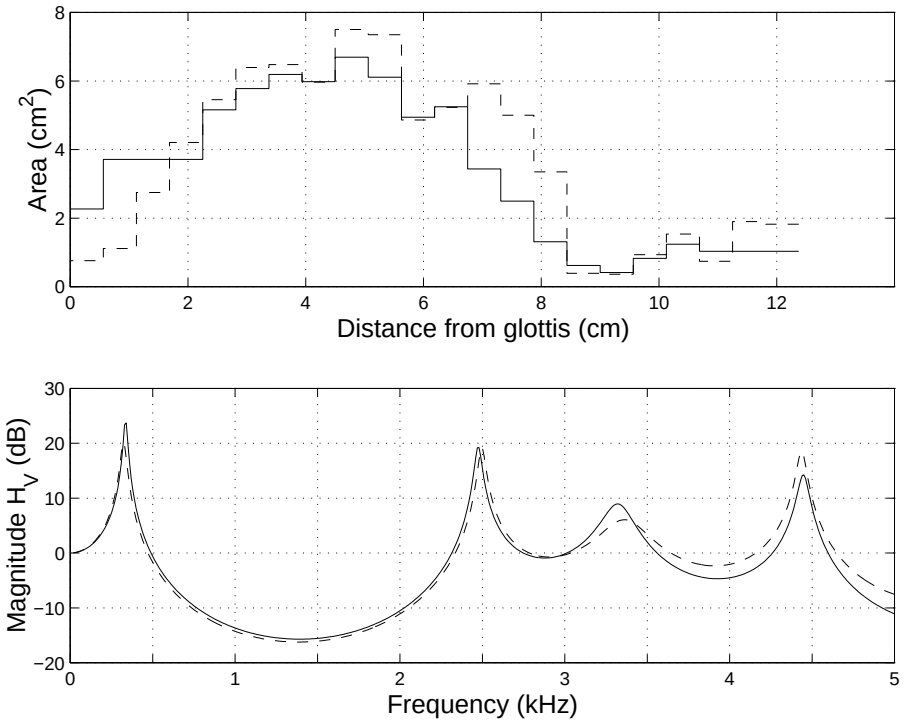


Figure 3.5: Geometry functions (top) from a single slice (frame 31, without lips or teeth), and the magnitude of transfer functions generated from them by VOAC (bottom). The solid line uses an assumption of circular cross section to calculate the area function, while the dashed line uses the method of Beautemps et al. (1996).

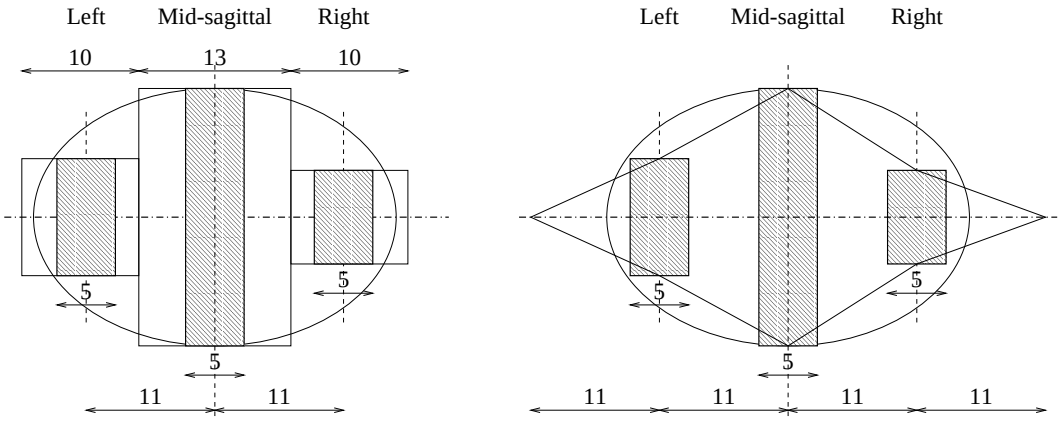


Figure 3.6: Slices combined as (left) blocks, and (right) polygon.

Fig. 3.6, left). Otherwise, for simplicity, the intervals might all be set equal to the inter-slice spacing of 11 mm, as per Figure 3.6 (right).

Termination at the glottis

Since the tract is considered closed at the glottis to a first approximation, differences in the location of the glottal end of the tract for the left, middle and right slices can be accommodated by simply summing the areas as far as the point furthest upstream in the mid-sagittal slice, which is then labelled the glottis. Lateral area sections interior to the glottis may be incorporated as additional area contributions, or in a fuller description using side branches. For this study, we defined the position of the glottis from the mid-sagittal slice and added contributions from the other slices as long as they were present. There were no examples of the side area functions extending beyond the glottis as defined by the mid-sagittal slice.

Termination at the lips

Specifying the termination of the vocal tract at its open end is a more complicated problem since it raises the question of where the mouth opening is, which is modelled acoustically as a piston radiating sound to the far field. One possible approach is to take the average or median of the end points for each of the slices. In this instance, shorter slices may have their area extrapolated by some means that reflects the increasing area of a bell-like curved aperture, and the longer slices truncated to leave an appropriate radiating surface at the lips. In this study, we took a simpler approach which involved defining the position of the termination in line with the end of the mid-sagittal slice, and then either truncating longer side slices or extending them at their final values.

3.4.2 Side branches

Although VOAC provides for the modelling of side branches, it may be preferable initially to amalgamate some or all of them with the main cavity for the sake of simplicity. The three levels of complexity are: (i) use only the main cavity, (ii) combine all branches to a single one, and (iii) model side branches individually. Combination of areas might be performed by summation of the side-branch areas with that of the main branch, and an area-weighted average for the hydraulic radius. In general, it is better to supervise the modelling of sinuses and to decide manually whether any side branch should be merged with the main cavity or not. Where large data sets render this approach impractical, rules need to be devised that take account of the size of the side branches and their anatomical location. For example, the pyriform sinuses, being unconnected with the main tract except at their aperture, should perhaps be modelled as a side branch; whereas the area behind the velum is well-coupled to the main tract, and

could be subsumed into it. Moreover, to add to the complexity of the problem, the sublingual cavity is sometimes translated by the outline conversion into a side branch within the distance function, meanwhile the pyriforms often appear as the main tract in the left and right slices, instead of as attached branches.

3.4.3 Area functions

Four area functions obtained by this technique are plotted in Figure 3.7, which correspond to the mid-phone frames of the nonsense word /pɑsi/ (for [p] this was taken as being just before release). In these examples, the side branches were discarded. The area functions for the vowels [ɑ] and [i] roughly approximate what might be expected for a low back vowel and a high front one, according to the tongue position. Yet for [ɑ], the area of the pharynx where the back of the tongue narrows the tract (4–7 cm from the glottis) was wide in comparison to measurements by Baer et al. (1991). Similarly, the constriction in the [s] area function has an atypically large area, of approximately $1\,\mathrm{cm}^2$, at a distance 14 cm from the glottis. Narayanan et al. (1995) report minimum constriction areas for [s] of 0.1–0.3 cm^2 for their four subjects. Their subjects sustained the fricatives, which would tend to result in smaller constrictions, but the discrepancy is still large.

The resolution of the dMRI images is 1 pixel within the plane of a slice, which corresponds for these images to 1.875 mm, and an area of $0.2\,\mathrm{cm}^2$. If each of the three slices has a sagittal distance from tongue to palate of one pixel, the minimum constriction area is therefore $0.6\,\mathrm{cm}^2$. In the mid-fricative frame, the minimum distance across the constriction in each slice was one pixel, but these points were not precisely the same length from the glottis. The rapidity of the area change for [s] also acts to decrease the dMRI resolution. Finally, the position of the teeth within the image was estimated manually by adjusting the brightness and contrast of the images and making a judgement, introducing additional uncertainty in the vicinity of the constriction. A point to note is that the position of the constriction is not consistent across the three sagittal slices, if one examines the outlines for [p] and [s] from within the reference frame of the overlaid grid (see Fig. C.3). This would suggest that a more sophisticated strategy is required for combining the slices in the anterior part of the mouth, where the shape of the palate differs considerably from left, through mid-sagittal, to right.

3.5 Computing VTTFs from real speech data

Having acquired a one-dimensional description of the vocal-tract, we are now in a position to model the acoustic properties of the duct. However, we must first encode the geometry functions in a way that VOAC can interpret.

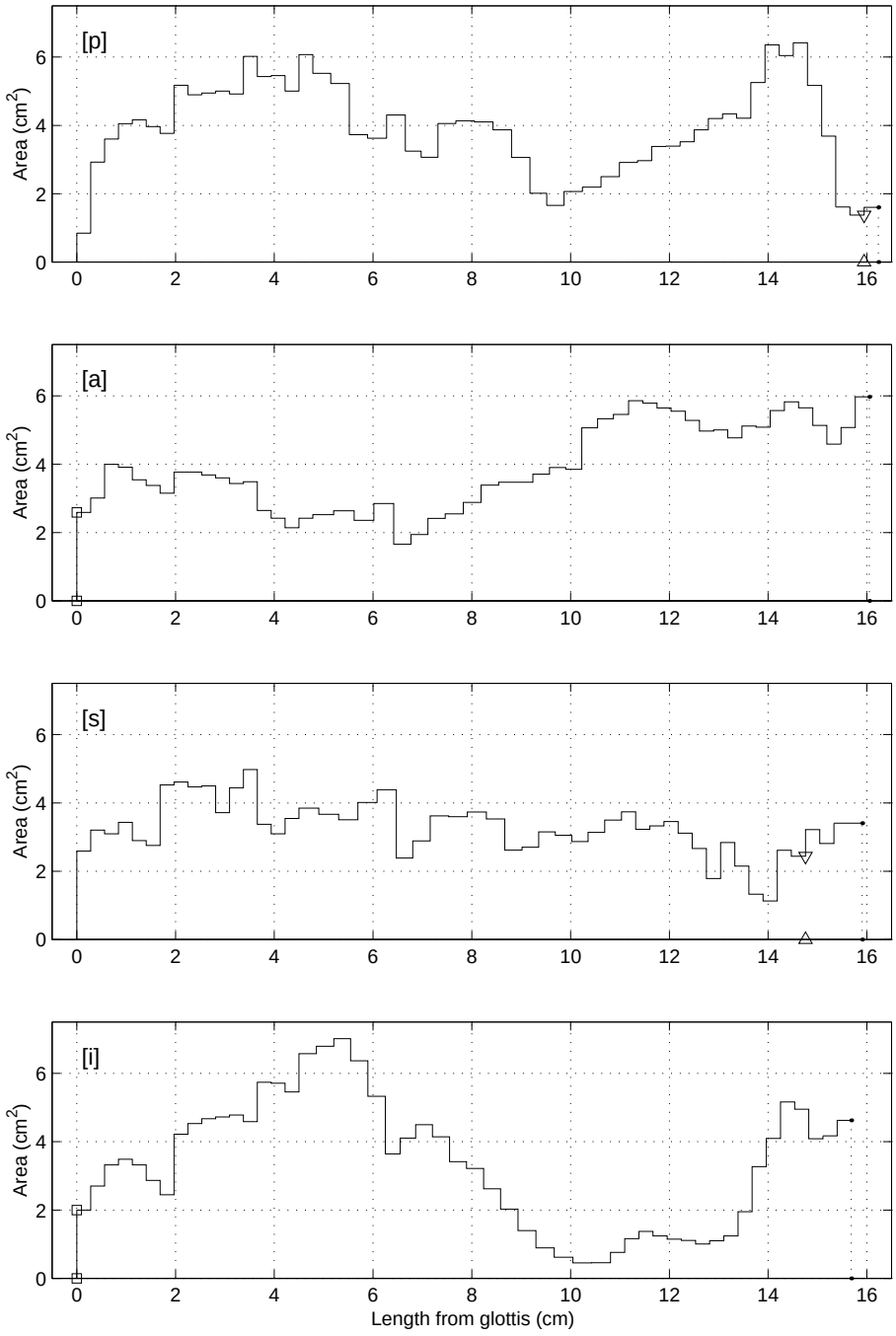


Figure 3.7: Area functions combining sagittal slices from the mid-points of four phones from the dynamic MRI data of [p$^{\text{h}}$ɑsi] by PJ: (from top) plosive [p], vowel [ɑ], fricative [s], and vowel [i]. The radiation surface is shown as a dotted line. For the vowels, the volume-velocity source was at the glottis, as indicated by the squares; for the consonants [p] and [s], the pressure source was at the lips and teeth respectively, as indicated by the triangles.

3.5.1 Generating input files for VOAC

No comprehensive set of rules has been written for selecting the discrete element types to represent the vocal-tract area function. For the following examples, all area changes were modelled by abrupt elements (ORIFICE and OUTLET). The geometry functions were converted into input files for VOAC minimally, by using only Type 1 for an expansion, Type 4 for a contraction and Type 5 for stretches of constant area. Thus, only the first one or two sub-elements of each type were used, leading to representations with a large number of elements (almost as many as there were gridlines). A more recent version of VOAC (v4.5) than the one we inherited uses an abruptness criterion for automatic element selection, testing the area gradient $\alpha = \delta S/\delta x$. If the area change is slight ($\alpha < 0.04$), then a RAMP element is used, in place of an abrupt contraction or expansion. As the number of elements representing the vocal tract, and hence the resolution, is increased, the placement of elements with large abrupt steps should become clearer and ultimately converge towards the actual distance profile specified by the vocal-tract outline. The alignment and representation of side branches is highly dependent on anatomical detail and can currently only be attempted manually by an acoustics expert familiar with the physiology of speech production. However, by absorbing side branches into the main tract and using abrupt contractions and expansions, a reasonable approximation to the true vocal-tract shape can be achieved, as demonstrated by the results that follow.

3.5.2 Vocal-tract transfer functions

The VTTFs calculated by VOAC for [p], [ɑ], [i] and [s] are given in Figure 3.8. The formants of the two vowels differed markedly. F1, F2, F3 and F4 were respectively 0.60, 1.34, 2.67, 3.58 kHz for [ɑ], and 0.36, 2.14, 2.49, 3.60 kHz for [i]. Formant frequencies extracted from speech recordings for the same subject (PJ) were 0.7, 1.1, 2.7, 3.6 kHz for [ɑ], and 0.3, 2.3, 2.9, 4.3 kHz for [i]. In comparison, therefore, the predicted values of F2, F3 and F4 for the vowel [i] were too low, whereas all except F2 matched well for [ɑ]. In relation to the resonances of a neutral vowel (i.e., 0.5, 1.5, 2.5 kHz, etc.), the predicted values were all nearer than the corresponding measured ones. Neither VTTF is as expected for [s], though they do correspond well to the area function in Fig. 3.7. The pressure VTTF for the fricative, $H^P_{QL}(f)$, shows the spectral zeros introduced by the rear-tract transfer function, which is a characteristic of localised supraglottal sources. The plosive's VTTFs are similar to those of [s] but with lower formants, as expected for a more anterior constriction. There is also less damping, yet the overall form of the pressure VTTF contains the peaks and valleys that are characteristic of the stop consonant. Note that the area functions from which these VTTFs were calculated were derived directly from the dMRI data, and no attempt has been made to modify them in relation to observations, as was done by, for instance, Beautemps et al. (1995).

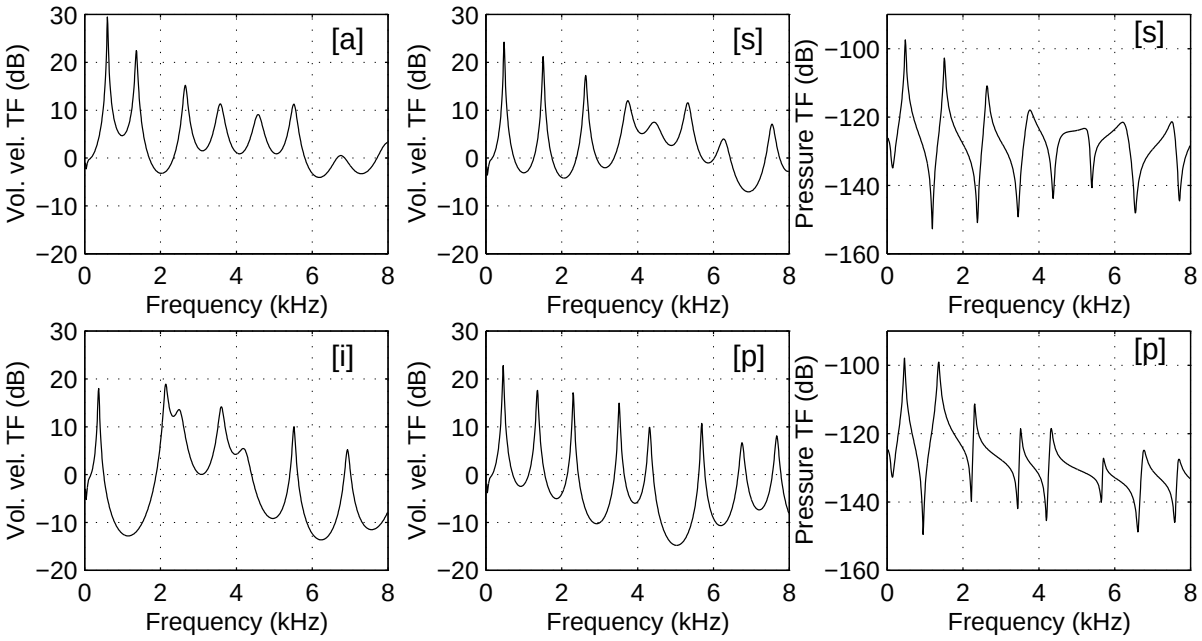


Figure 3.8: Transfer functions predicted by VOAC for the four phones, [p, ɑ, s, i]. The VTTFs for a volume-velocity source $H_{GL}^{V}(f)$ at the glottis are given for the vowels (left), [ɑ] (top) and [i] (bottom); for the consonants (centre), [s] (top) and [p] (bottom). The VTTFs for a pressure source $H_{QL}^{P}(f)$ are also given (right) for the consonants [s] (top) and [p] (bottom), for which the source was located downstream of the constriction for the fricative [s] and at the lips for [p].

3.6 Speech synthesis

As a further extension of the predictions, the VTTFs were used to synthesise some speech-like sounds, which could be used as a further assessment of the modelling and acquisition procedures.

3.6.1 Overview

The principal issues to be addressed in the synthesis of speech are as follows: source type (pressure/volume velocity), source impedance/admittance, and source location, which in part determine the characteristics of the filter. An intermediate source is generally not located at the constriction. It should be placed a short distance downstream, or at an obstacle, such as the teeth. With the inclusion of a source impedance, the pressure source can alternatively be represented as a volume-velocity source. There is an equivalence between the two source types, akin to the current-voltage equivalence (Norton/Thevenin) in electrical circuits.

For some sounds, the position of the acoustic source is more obvious than for others. For instance, /ʃ/ quite clearly produces a source at the teeth, whereas /x/ has sources that are physically distributed along the roof of the palate (Shadle 1991). In a study by Narayanan and

Alwan (1996), the source types were broken down into flow monopoles, dipoles and quadrupoles. Not only did they identify differing source locations for fricatives of different place, but they found that the best results were obtained by placing the components at differing points, such as, for [s], a dipole at the teeth and a monopole at the constriction exit.

For our synthesis of the vowels /ɑ/ and /i/, we have assumed that there is a volume velocity source at the glottis with a hypothetical waveform. For the fricative /s/, a pressure source of coloured noise was located a short distance (1 cm) downstream of the constriction exit. The release of the stop consonant /p/ was excited by a purely transient pressure signal injected just inside the obstruction at the lips, and no attempt was made, at this stage, to include the subsequent fricative and aspirative contributions.

3.6.2 Impulse response filter

For a volume-velocity source at the glottis U_G, such as voicing, the volume velocity at the lips was calculated from the volume-velocity VTTF:

$$H^V_{GL}(f) = \frac{U_L(f)}{U_G(f)}\,. \tag{3.7}$$

The radiation from the lips was approximated by a piston in an infinite baffle (Beranek 1954), which was used to determine the terminal reflection coefficient. To predict the far-field sound p_{ff} radiated from U_L at $r = 0.3\,\mathrm{m}$, the VTTFs were multiplied by the radiation factor $\rho f/r$, where ρ is the density of air, according to Eq. 2.4.

Modelled as an ideal pressure source within the tract, the frication source p_Q induced waves travelling both upstream (towards the glottis) and downstream (towards the lips). As explained in Section 2.4.2, the overall VTTF from the source to the lips is equal to the product of two transfer functions:

$$H^P_{QL}(f) = \frac{U_G(f)}{p_Q(f)}\,\frac{U_L(f)}{U_G(f)} = H^P_{QG}(f)\,H^V_{GL}(f)\,, \tag{3.8}$$

where H^P_{QG} is the pressure transfer function of the rear-tract, that part upstream of the source, which uses a reflection coefficient $R = 1$ at Q. To be able to compute $H^P_{QG}\big|_{R=1}$ using VOAC, we provided it a flag, which was set `TRUE` for the unit reflection coefficient (i.e., infinite impedance), and otherwise `FALSE` (i.e., for piston-in-baffle radiation impedance). Thus, supplying a geometry function defined only from the glottis to the source location, VOAC computed the transfer function from a volume-velocity at G to a pressure at Q:

$$\left(\frac{p_Q(f)}{U_G(f)}\right)_{R=1} = \frac{1}{H^P_{QG}(f)}\,, \tag{3.9}$$

from which the desired rear-tract transfer function can be obtained by the principle of reciprocity.

The VTTF was usually computed at 10 Hz intervals from the specified Nyquist frequency, i.e., 8 kHz for sample rate $f_s = 16\,\mathrm{kHz}$, down to 20 Hz, since the frequency domain algorithms

are invalid close to zero frequency. The VTTF was then extrapolated down to d.c. to provide a complete spectrum, ready for inverse transformation. For a volume-velocity source, such as voicing, the volume velocity flowing into the vocal tract is the same as that leaving it, at very low frequencies. So, the zero-frequency response was set to unity (0 dB) and the intervening point was the geometric mean of its neighbours:

$$H^V(0) = 1, \tag{3.10}$$

$$H^V(10) = \sqrt{H^V(20)}. \tag{3.11}$$

Considering the rear-tract only, the reflection coefficient $R = 1$ at the source location implies that the slightest volume velocity (current) will induce a pressure (potential difference) at the source plane (the terminals of the open circuit). Since there is effectively no resistance at very low frequencies, the pressure VTTF becomes infinite and hence the overall pressure to lip volume-velocity VTTF tends to zero. For the pressure source, the zero-frequency response was set to zero, and the intervening point was the arithmetic mean:

$$H^P(0) = 0, \tag{3.12}$$

$$H^P(10) = \frac{1}{2}H^P(20). \tag{3.13}$$

To obtain the real impulse response functions at the desired sample rate, each extrapolated, radiation-adjusted VTTF was appended with its complex conjugate mirror image. The VTTF could be zero-padded in the upper frequency region, to match a higher sampling rate if required. Finally, the whole array was inverse Fourier transformed to yield the predicted impulse response of the vocal tract to the specified source.

3.6.3 Acoustic sources

Voicing

A very simple glottal source model was used to excite the vocal tract for synthesising the voiced component. No perturbation in pitch or amplitude was added, but f_0, which was centred on 131 Hz, declined linearly throughout the synthetic phone by approximately 5 Hz.

The waveform $g(n)$ was comprised of discrete "open" and "closed" phases, which were constructed piecewise from a cubic function and a constant amplitude section, respectively. The open quotient, the ratio of the open portion to the total pitch period, was fixed at OQ = 0.5, and the amplitude at unity. The amplitudes of the closed flow U_{closed} and the open flow U_{open} were used to define the overall gain of the signal, giving:

$$g(n) = \begin{cases} U_{\text{closed}} + U_{\text{open}}\left(a_0 + a_1 n + a_2 n^2 + a_3 n^3\right) & \text{for } 1 \le n < 0.5T_0, \\ U_{\text{closed}} & \text{for } 0.5T_0 < n \le T_0. \end{cases} \tag{3.14}$$

The cubic coefficients were $a_i = \left\{0,\ 0,\ 27/4\,T_{\text{open}}^2,\ -27/4\,T_{\text{open}}^3\right\}$, where $T_{\text{open}} = \text{OQ}\,T_0$. The idealised glottal waveform, shown in Figure 3.9, was generated by the cubic formula, Eq. 3.14, using $U_{\text{open}} = 0.3$ and $U_{\text{closed}} = 0.1$.

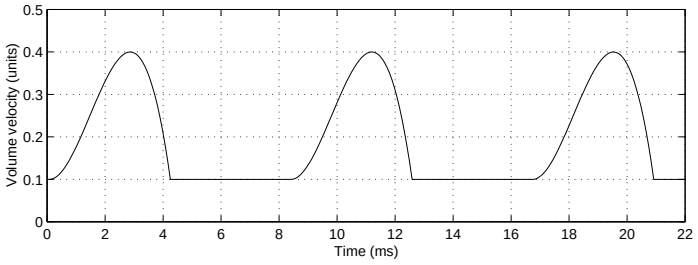


Figure 3.9: Glottal source waveform at a constant fundamental frequency, $f_0 = 120\,\text{Hz}$, with a cubic profile during the open phase lasting one half of the cycle.

Noise

The transient source was a localised pressure pulse with an exponential decay:

$$d(n) = \begin{cases} 0 & \text{for } n \leq 0, \\ \exp\left(-an\right) & \text{for } n > 0, \end{cases} \tag{3.15}$$

where a was calculated to give a half-life equal to 5 ms.

The frication-noise source was generated from Gaussian white noise (provided by a pseudo-random number generator, using the Matlab function `randn`) that was coloured to reflect the required source characteristics. Thus, to synthesise the unvoiced fricative /s/, a white noise spectrum $N(k)$ was coloured, according to Shadle's (1985) estimated regression curve:

$$D(k) = N(k)\,a \exp\left(\frac{bkf_s}{N}\right), \tag{3.16}$$

where the constants are $a = 95$ and $b = -0.0004$, N is the total number of points, f_s is the sampling frequency, and $D(k)$ is the source spectrum. The time series $d(n)$ was obtained for the source signal by computing the inverse Fourier transform of $D(k)$.

3.6.4 Results

Each acoustic source signal was convolved with the impulse response function obtained from the appropriate VTTF including the effects of radiation, and the overall time envelope of each signal was adjusted to give slightly more natural sounding onset and offset characteristics. The results of the synthesis procedure are available from the project web-site (Jackson 1998).

The vowels had a highly artificial sound with an almost metallic timbre, yet with F1 and F2 similar to those observed from recordings of the same subject, PJ, they were clearly recognizable as approximations of the intended phonemes. The high /i/ sounded correct, but the open /ɑ/

vowel tended to sound more neutral than expected, approaching a [ə]. The reason for this is most likely to be the result of inaccuracies in the estimation of the area functions, as noted earlier in Section 3.4.3.

The sound of the burst phase from the plosive /p/ was reminiscent of the opening of a jam jar, since no frication or aspiration phases were adjoined. In fact, the synthetic sound can quite easily be produced by a human speaker, when the glottis is closed and no substantial air flow follows after the release to drown the burst sound with frication noise.

The fricative /s/ produced a sound rather like a static flow duct noise. Unlike a real [s], the sound came mainly from low frequencies and was dominated by F1 and F2. This effect depends, in part, on the position of the sound source in relation to the constriction, but mostly on the size of the constriction. With the constriction being too big, there was not enough damping of the rear tract. We have chosen to synthesise the fricative with the estimated area function as drawn in Figure 3.7, rather than to attempt to correct it at this stage. Later, in Chapter 7, we use this area function to synthesise /z/, as well as the /s/ here. Although large, we have not decreased the area in the constriction simply because we expect it to be smaller and know that it is an acoustically sensitive dimension. Further tests reducing the constriction size to $0.2\,\mathrm{cm}^2$ yielded dramatic low-frequency attenuation, reducing F1 and F2 by nearly 20 dB.

3.7 Summary

In this chapter, we have extracted area and hydraulic radius functions from dMRI data and used the geometrical information contained in the images to synthesise speech-like signals. Our interpretation of the dMRI vocal-tract outlines resulted in reasonable area functions for [p], [ɑ], [s], and [i], although we lacked resolution for the regions near the plosive and fricative constrictions, and the area function for [ɑ] tended towards that of [ə]. Differences in these regions may also be attributed to the dynamic context in which the consonants were produced. Future attempts might seek to resolve this deficiency by incorporating data from electropalatography, static MRI, or even video, in the case of labials.

Further work is needed to provide a complete dynamic sequence from the dMRI frames, but some aspects of the conversion to geometry functions need to be improved, either by closer supervision of the distance function processing or by using higher resolution images. It may be that the sagittal spacing of the dMRI slices was rather too wide to provide the sort of resolution required for accurate representation of the vocal tract along one dimension. Nevertheless, speech-like sounds were successfully rendered using the simulated VTTFs and assumed source profiles. Although the quality of the synthetic signals was unnatural in many respects, this procedure demonstrated the capability to produce speech sounds from medical

images in a way that could be incorporated into an articulatory synthesiser. Moreover, certain aspects of the artificial sounds, such as the vowels' formants and plosive burst's anti-resonances, were characteristic of the phonemes they represented. Also, there is the capability of performing comparisons with analyses of frication and aspiration, as a tool for speech production studies, which we will further expound in the succeeding chapters.

Chapter 4

Analysis of single-source speech

This chapter describes a range of methods for analysing a speech signal as a single, complete entity. First, details are given of a collection of speech recordings that were designed to provide suitable material for the analyses selected for the present study. Related issues such as ways of improving the measurement quality for an adaptive model are considered, and some difficulties that arise when analysing real speech. Results are presented of applying these techniques to unvoiced plosives, which are mainly excited by a single source at any one time, although a number of source mechanisms is in operation.

4.1 Speech acquisition

The first step in performing an analysis of speech materials is to make suitable recordings. This section describes a series of six recording sessions, with progressively more elaborate corpora and instrumentation.

4.1.1 Subjects

To obtain the necessary data for analysis, speech recordings were made from a pool of four subjects, two male and two female. They were all healthy adults who had no known speech pathologies: male PJ, 27–28 years old, native speaker of British English (received pronunciation); female CS, mid 40s, American English (California); male LJ, mid 20s, European Portuguese; female SB, late 20s, British English (Lincolnshire).

4.1.2 Corpora

Corpus 1

As a preliminary exercise to obtain examples of real speech, some pilot recordings were made, using a microphone held at approximately 0.1 m from the lips and the CSL Kay system. The

sound pressure was acquired in the laboratory at 10 kHz sampling rate, with digital (equi-ripple FIR) anti-alias filtering rolling off at 4 kHz by 400 dB/decade. The corpus, to which we will refer as Corpus 1 or *C1*, contained a series of /CV/-syllables spoken by subject PJ.

Corpus 2

A second set of recordings, *C2*, was made by PJ in a sound-treated booth to reduce the level of ambient noise. The corpus comprised repetitions of /pɑ/ in three modes of phonation: modal, pressed (or stage-whispered) and whispered. For each mode, at least eight valid records were obtained, covering a range of pitches and intonation patterns.

The sampling rate was increased relative to *C1* to reduce time-quantisation ($f_s = 48$ kHz), and to provide information up to 20 kHz. The acquisition procedure had two steps: (i) the sound pressure was measured by a microphone (Brüel and Kjær Type 4133, via a B&K 2639 pre-amplifier and B&K 2636 amplifier with 22.4 Hz–22.4 kHz band-pass, linear filtering) at 0.3 m from the lips directly in front of the subject, and was stored with a DAT recorder (Sony TCD-D7), and (ii) the recording was later replayed from tape and digitally transferred to computer as 16-bit data at 48 kHz. A calibration tone was recorded to give an absolute reference to pressure (B&K 4230 provided a 93.8 dB re. 2×10^{-5} Pa at 1 kHz), and background noise was recorded to allow assessment of the measurement-error noise floor.

Corpus 3

Corpus *C3* consisted of repeated /CV_1FV_2/ nonsense words with ten repetitions in one breath by PJ. The consonants C were unvoiced plosives and were included as an important instance of aspiration noise, and fricatives F were included as an alternative source of turbulence noise (to plosion and aspiration). In the case of voiced fricatives, the vowels provided a stable voicing context across the two syllables of the utterance. Different voicing qualities were achieved by varying the utterance's mode of phonation, which among other things served to modify the balance of voicing and aspiration noise.

The vowels V_1=V_2=/ɑ, i, u/ were chosen since they are native vowels that have quite separate articulatory configurations and exercise much of the vowel space: low, high-front and high-back, respectively. The unvoiced stop consonants C=/p, t, k/ were chosen since they tend to produce more aspiration noise than the voiced ones, and represent the only three places of articulation in English stops: labial, alveolar and velar, respectively. The fricatives F=/s, z/ were chosen to give a voicing contrast for the most common example, enabling direct comparison of related single-source and mixed-source phonemes. Various phonation modes were employed: modal, breathy, pressed and whispered. Sustained fricatives F=/s, z/ were also recorded in the context /ɑF:/ to provide a stable condition for examination under other

analysis techniques, e.g., time-averaging. Not all combinations were recorded, but enough to allow a full set of comparisons: effect of vowel context, e.g., /pɑsɑ, pisi, pusu/; effect of place, e.g., /pɑsɑ, tɑsɑ, kɑsɑ/; effect of phonation mode, e.g., /pɑsɑ/ modal, breathy, pressed and whispered; effect of voicing in a fricative, /pɑsɑ, pɑzɑ/; effect of duration, e.g., /pɑzɑ, ɑzː/.

The recording procedure was identical to that of *C2*, but in stereo. The sound pressures at 30° azimuth and 0.5 m, and straight ahead at 1 m (0° azimuth, both 0° elevation) were measured using B&K 4133 and 4165 microphones, B&K 2639 pre-amplifiers, and B&K 2609 and 2636 measurement amplifiers, respectively.

Corpus 4

Two subjects, one male (LJ) and one female (SB), recorded the speech corpus *C4*, which contained sustained vowels V = /ɑ, i, u/ and fricatives /ɑsː, ɑzː/, and nonsense words, /hɑhɑ/ in a carrier phrase and /pɑzɑ/. The recording procedure was identical to that of *C2* with the microphone (B&K 4133) at 1 m, except that an electroglottograph (EGG) was simultaneously recorded onto the second DAT channel. The EGG (Laryngograph Lx Proc PCLX) was used to measure the transglottal impedance with adult (large) electrodes, and its phase response was checked using a square-wave input signal.

Corpus 5

An extended version of *C4* was recorded by PJ. This corpus, *C5*, contained sustained fricatives (not all of them native to English), sustained vowels and three kinds of nonsense word: /pɑsi/, /pɑFɑ/ where F ∈ {ɸ, β, f, v, θ, ð, s, z, ʃ, ʒ, x, ɣ, ħ, ʕ}, and /CiFi/ using C-F pairs /p-ɸ, b-β, t-s, d-z, k-ç, g-j/. These words were repeated to give 10 tokens using a single breath. All the fricatives were sustained for 5 s in /ɑFː/ context, and a subset /ɸ, β, s, z, x, ɣ/ were spoken loudly and softly, and with increasing amplitude. Others /f, v, θ, ð/ were uttered in a different vowel context /iFː/. The vowels /ɑ, i, u/ were spoken with differing voice qualities, i.e., modal, pressed, breathy and whispered, and were also sustained for 5 s. The recording procedure was otherwise identical to that of *C4*.

Corpus 6

The corpus *C6* consisted of sustained voiced fricatives F = /v, ð, z, ʒ/ recorded in /əFː/ context by PJ and CS. The purpose was to capture sustained fricatives that were adjusted in a minimal sense by a change in pitch, place or mode of phonation. For reference, there were fricatives sustained at constant pitch, in addition to both ascending and descending f_0 glides. Others captured transitions of place within a phoneme /əzː, əʒː/ and between phonemes /əzːʒː, əʒzː/,

and from voiced to unvoiced /əzːsː, əʒːʃː/. The recording procedure was otherwise identical to that of *C4*.

4.2 Analysis in the frequency domain

Transformations of the signal into other domains present an opportunity to identify features that may be a dominant characteristic in the new domain. The mechanics of the inner ear act in many respects as a continuous array of band-pass filters that separate frequencies spatially along the cochlea, before being transmitted down the auditory nerve. Therefore, it is logical to apply some kind of frequency transformation on speech signals, since we expect that much of the signals' salience will emerge. (Later, in Chapter 7, we will investigate a form of time-series analysis that seems to capture another aspect of perceptual sensitivity.) The Fourier transform provides a precise and unique frequency transformation that is fully reversible, i.e., with no loss of information:

$$S(j\omega) = \int_{-\infty}^{\infty} s(t) \exp(-j\omega t)\, dt\,, \tag{4.1}$$

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(j\omega) \exp(j\omega t)\, d\omega\,. \tag{4.2}$$

It assumes that signals are made up of a superposition of sinusoids that are infinite in extent over time. In reality, a finite section of speech is analysed at any one time, which has been sampled at discrete intervals. The discrete Fourier transform pair can be written thus,

$$S(k) = \sum_{n=0}^{N-1} s(n) \exp\left(-j\frac{2\pi nk}{N}\right), \tag{4.3}$$

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) \exp\left(j\frac{2\pi nk}{N}\right). \tag{4.4}$$

The resultant spectra contain complex coefficients, which are symmetric across frequency in magnitude, and anti-symmetric in phase, because they were obtained from real signals, i.e., they occur in complex conjugate pairs. They can be visualised by plotting their magnitude and phase separately up to the Nyquist or folding frequency, which is half the sample rate, $f_s/2$. Thus the sampling frequency determines the frequency range of the spectrum. The coarseness of the frequency resolution depends on the size of the section of speech analysed: the more points in the section, the more in the spectrum. This implies the immutable compromise between time resolution and frequency resolution. The discrete frequency values are often referred to as bins since, roughly speaking, the signal power over the band of frequencies corresponding to an individual bin are all collected by the bin, and so the wider the bin, the more power it gathers.

4.2.1 Windowing

Windowing enables finite time resolution to be achieved with the Fourier transform, at the inevitable expense of some spectral sharpness, and produces a result called the short-time Fourier transform (STFT). The pay-off between time and frequency resolution depends on the rate of change of features and the accuracy with which they are required, but the bandwidth-time product cannot be less than one half: $\Delta f \, \Delta t \geq \frac{1}{2}$. Nevertheless, some gains in compactness can be made using specific windows to shape the speech section before processing. Generally speaking, smooth functions whose derivatives are also smooth reduce the amount of spectral leakage or smearing. The main purpose of windowing therefore is to enable a small part of a signal to be processed such that it gives a response that is compact in both time and frequency.

Merely restricting the number of points selected, which is equivalent to applying a rectangular (or boxcar) window, results in sizeable sidelobes in the frequency domain, for any significant component, that manifest themselves as a considerable amount of smearing. Application of a continuous function, such as a triangular (or Bartlett) window, which only has discontinuities in its (first) derivative, yields a result with less spectral leakage. As the smoothness order increases, the leakage decreases, but with diminishing improvements. Other popular windows include the Blackman, Hann and Hamming functions, of which our preferred choice is the Hann window:

$$w(n) = 0.5 \, (1 - \cos 2\pi n/N) \text{ for } n \in \{0, 1, \ldots, N-1\}. \tag{4.5}$$

Windowing modifies the interpretation of the short-time Fourier transform from one of piecewise stationarity, to an adaptive quasi-stationary approximation of a dynamic system. Thus, using a smooth window function offers benefits in modelling for the following scenarios:

1. variations in the formants (centre frequency and bandwidth);
2. perturbations and transients in fundamental frequency f_0 (e.g., from jitter and at voice onset);
3. linear changes in f_0 (i.e., $df_0/dt \neq 0$);
4. higher order changes in f_0 (i.e., $d^n f_0/dt^n \neq 0$);
5. perturbations and transients in amplitude A (e.g., from shimmer and at voice onset);
6. linear changes in A (i.e., $dA/dt \neq 0$);
7. higher order changes in A (i.e., $d^n A/dt^n \neq 0$);
8. variations in other source characteristics (e.g., spectral colouring);

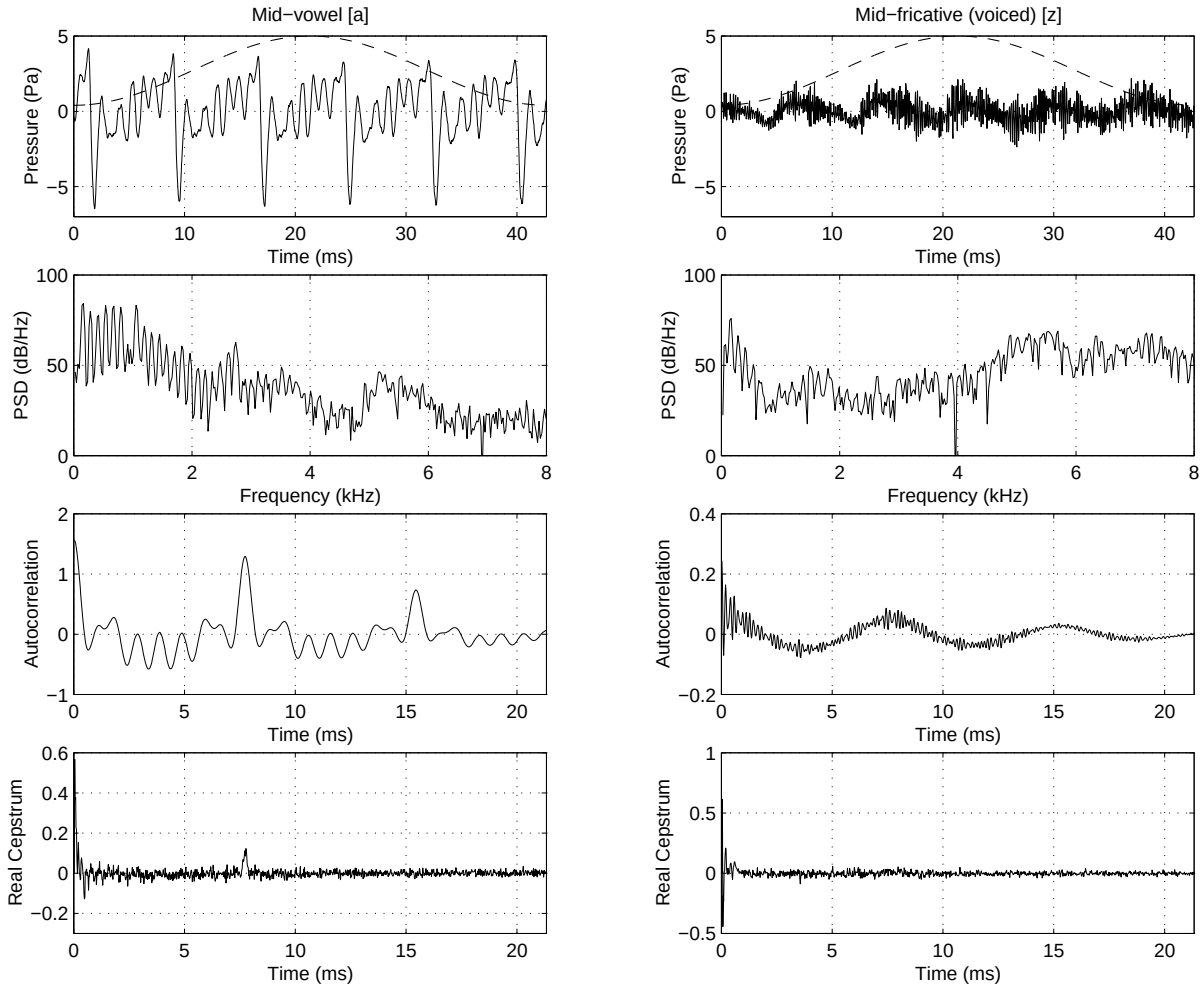


Figure 4.1: Four signal quantities illustrated for mid-phone examples of [ɑ] (left) and [z] (right): (from top) sound pressure x (solid) overlaid with the analysis window (dashed, 2048-point Hamming, c. 43 ms), power spectral density S_{xx}, the autocorrelation R_{xx}, and the real cepstrum C_x.

9. changing noise contributions from other acoustic sources;

10. disturbance, interference and measurement noise.

4.2.2 Power spectra and spectrograms

Since human perception of sound is approximately logarithmic for most of the ear's dynamic range, and the relative levels of different frequency components can vary dramatically, the magnitude sound spectrum is normally plotted on a decibel scale, normalised on the spectral bin width, and called the *power spectrum*. Thus, the area under the curve gives the contribution to the sound power from each frequency band, and the total area gives the SPL.

The power spectra for two phonemes are shown in Figure 4.1, plotted underneath their time series. For the vowel [ɑ] (left), the regular glottal pulsing, so clearly evident in the time signal, is evident in the power spectrum as harmonics of $f_0 \approx 130\,\text{Hz}$ up to about 4 kHz;

above 4 kHz, they disintegrate into a noise-like spectrum. Some resonances, which are labelled formants and manifested as broader spectral peaks in the envelope of the harmonics, can be seen, for example at 1.0 kHz, 2.7 kHz and 5.2 kHz, and an anti-resonance at 4.8 kHz, manifested as a spectral trough. In contrast, the signal of the voiced fricative [z] (right) is very noisy in appearance and its spectrum shows little harmonic structure. It has broad formants, such as at 1.5 kHz, 3.6 kHz and higher, and an anti-resonance, a low-frequency system zero, at 0.8 kHz. Below that, the weak effect of voicing stands out as a row of harmonics that rapidly disappear into the noise floor. The peaks 5–6 kHz and 6–8 kHz are very broad, making it difficult to distinguish individual resonances. Now, if we ignore the low-frequency energy from voicing and consider the overall slope of the spectrum, we see a rise of about 30 dB for the fricative over the range 1–8 kHz, as opposed to a fall of about 60 dB for the vowel. The spectral slope, or spectral tilt as it is known, is very different for these different classes of phoneme. Even within a single phonetic category, the spectral tilt can vary depending, for fricatives, on the turbulence-noise source strength and the constriction location (labiodental, alveolar, palatal, velar, etc.) and, for vowels, on the mode of phonation (e.g., modal, breathy, pressed). The energy in the first two or three harmonics is also a strong indication of the voice quality: it only accounts for about a third of the periodic signal power in [a], but almost all of the discernible harmonic power in [z].

By incrementing the window location gradually along the signal and computing the spectrum at each step, a picture can be built up of how the spectral characteristics develop over time. The spectrogram is constructed in just such a way with time plotted along the horizontal axis, frequency vertically, and the power spectral density represented by the grey level, on a decibel scale (or sometimes by colour). A waterfall plot is an alternative spectrographic means of representing the time variation of short-time spectra, which overlays the spectra at successive time instants with a small vertical offset. It visually accentuates changes over time in the spectral characteristics, both peaks and troughs.

4.2.3 Time-averaging

Each spectrum that is computed suffers degradation from interfering noise. When a steady sound is produced that is sustained over many tens of milliseconds, a number of measurements can be made by placing successive analysis frames along the signal. If these frames are each weighted by a window function and placed end-to-end, the parts of the signal at the frame boundaries will receive less prominence in the analysis than those in the centre of a window. To make better use of the available data, windows may be overlapped although the gains are negligible beyond 50 % overlap, because of the degree of redundancy created.

For random variable x with a probability distribution function $f(x)$, the discrete and con-

tinuous means are written:

$$\mu_d = \sum_i x_i f(x_i), \tag{4.6}$$

$$\mu_c = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{4.7}$$

and the corresponding variances,

$$\sigma_d^2 = \sum_i (x_i - \mu_d)^2 f(x_i), \tag{4.8}$$

$$\sigma_c^2 = \int_{-\infty}^{\infty} (x - \mu_c)^2 f(x)\, dx. \tag{4.9}$$

For statistically sampled variables, the mean and (unbiased) variance are:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{4.10}$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2 . \tag{4.11}$$

For repeated measurements of the same quantities,

$$\begin{aligned} \mathrm{E}[\bar{x}] &= \frac{\mu_1 + \mu_2 + \ldots + \mu_N}{N} \\ &\to \mu, \qquad \text{as } N \to \infty, \end{aligned} \tag{4.12}$$

$$\begin{aligned} \mathrm{E}[\sigma^2] &= \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_N^2}{N^2} \\ &\to \frac{\sigma^2}{N}, \qquad \text{as } N \to \infty, \end{aligned} \tag{4.13}$$

which implies that the standard deviation of an estimate can be reduced by a factor of $\sqrt{N}$ by averaging over N measurements:

$$\bar{S}^2(f) = \frac{1}{N}\sum_{i=1}^{N} \hat{S}_i^2(f), \tag{4.14}$$

where $\hat{S}_i(f)$ is the spectral estimate obtained from each frame. If eight similar sections of the speech recording were used, a reduction of 9 dB ($20 \log \sqrt{8} = 9$) would be expected on the deviation, or "hair", of the spectrum. Also, we know that the frequency resolution can be increased by concatenating the records, as described above.

	binwidth	Deterministic		Stochastic	
frame(s)		power	var.	power	var.
1 single	1	1	1	1	1
1 double	1/2	2/2	1/2	1/2	1/2
N single	1	1	$1/N$	1	$1/N$

Table 4.1: Summary of mean power and power variance for power spectral estimates.

Let us consider averaging the power spectra of two kinds of signal: a deterministic signal and a stochastic one, and let us assume that the deterministic signal is a sinusoid of constant amplitude and frequency, and that the stochastic signal is comprised of uncorrelated random noise. Table 4.1 summarises the influence on variance of the power spectrum of the two rival ways to take advantage of additional data: lengthening the analysis frame for improving frequency resolution, and averaging across frames to increase the measurement accuracy (Bendat and Piersol 1984, pp. 191–192).

4.2.4 Ensemble averaging

Averaging is the practice of combining measurements of several instances of the same event to gain a more reliable description of that event. In a sustained sound, these instances effectively occur contiguously, but if the event is transient or dynamic in nature, the battery of tokens must be acquired by repeating the same articulatory sequence that produced the event. Once an ensemble of repetitions has been recorded, care must be taken with alignment, so that the event occurs at the same point in the analysis frame, for each of the tokens. The aligned tokens thus provide a synchronised array of events, which may then be combined in an ensemble average. For example, sections of speech from reiterations of a plosive-vowel syllable might be aligned on the stop release, to capture the characteristics of the burst event. In this case, as in analyses that are synchronised to individual glottal pulses (i.e., pitch synchronously), the phase of the tokens is also matched in the alignment and the time signals themselves may be averaged — equivalent to a direct average of the complex Fourier coefficients in the frequency spectrum. In /VFV/ sequences where F is a fricative, there may be no precise event to which to synchronise, not even glottal pulses for an unvoiced fricative. However, there may be a region, perhaps mid-phoneme, where the essential features of the frication are most pronounced. Here, it is meaningless to average either the time signals or the frequency spectrum, but averaging the power spectrum, on the other hand, can provide a means to capture and enhance the features that we want to quantify. Crucially, the error of the averaged spectrum is decreased with respect to that of each token.

4.3 Fundamental frequency

As noted, voicing is a dominant aspect of many speech sounds. One of its principal characteristics is the fundamental frequency of oscillation of the vocal folds. Some measures of the quality of voicing that are derived from f_0 are listed, and then we discuss a variety of approaches to automatic pitch extraction.

4.3.1 Perturbation measures

Jitter

Jitter is a measure of fluctuation in the pitch period (or conversely the fundamental frequency) of the voice.[1] Usually expressed as a percentage, it is defined (Horii 1979; Hillenbrand 1987; Dejonckere and Lebacq 1996) as:

$$\hat{\sigma}_T = \frac{\mathrm{E}\left[|\tau_i - \tau_{i-1}|\right]}{\mathrm{E}\left[\tau_i\right]} \times 100 \ (\%)\,, \tag{4.15}$$

where the period of the ith pulse, $\tau_i = t_i - t_{i-1}$, is the difference between the current pitch instant t_i and the previous one, and $\mathrm{E}[\,]$ denotes the expected value. It can be evaluated for all pulses in a given section of signal, or restricted to a region of that signal, to give a more time-specific measurement.

When analysing real speech, the jitter and equilibrium fundamental frequency vary with time. So, using a window function $x(n)$, e.g., Bartlett, Blackman, Hann or Hamming, offers a means to evaluate the jitter locally:

$$\tilde{\sigma}_T(p) = \frac{\langle |\tau_i - \tau_{i-1}| \ x(t_i - p) \rangle}{\langle \tau_i \, x(t_i - p) \rangle} \times 100 \ (\%)\,, \tag{4.16}$$

in the vicinity of point p, where $\langle\ \rangle$ denotes the time average. Note that, in practice, computation of Eq. 4.15 over a finite number of pitch periods is equivalent to Eq. 4.16, when $x(n)$ is rectangular (aka. boxcar). To identify the pitch instants, T_i, we supervised a combination of zero-crossing (Awan and Frenkel 1994) and peak-picking (Howard and Fourcin 1983) to enhance manual estimates.

Shimmer

Shimmer is a measure of the fluctuation of the amplitude of the voice. Usually expressed in decibels, it is defined (Hillenbrand 1987; Blomgren et al. 1998) as:

$$\hat{\sigma}_A = 20 \log_{10} \left(\frac{\mathrm{E}\left[|a_i - a_{i-1}|\right]}{\mathrm{E}\left[a_i\right]} \right) \ (\mathrm{dB})\,, \tag{4.17}$$

where a_i is the amplitude of the ith pulse. The corresponding windowed shimmer was:

$$\tilde{\sigma}_A(p) = 20 \log_{10} \left(\frac{\langle |a_i - a_{i-1}| \ x(t_i - p) \rangle}{\langle a_i \, x(t_i - p) \rangle} \right) \ (\mathrm{dB})\,. \tag{4.18}$$

For real speech, each pulse amplitude, a_i, was estimated using the RMS amplitude of the signal, windowed by an asymmetric Hann window, extending one pitch period either side of the pitch instant in question.

[1] Although pitch strictly refers to the perceptual effect of a certain fundamental frequency or f_0, we will use pitch as an adjectival noun referring to f_0, since the distinction is not of relevance to the present study.

Harmonics-to-noise ratio

The harmonics-to-noise ratio (HNR) is often used as a measure of the relative amplitudes of the voiced and unvoiced components. It is defined (Lim et al. 1978; Hillenbrand 1987) as:

$$\hat{\sigma}_N = 10 \log_{10} \left(\frac{\mathrm{E}\left[v^2\right]}{\mathrm{E}\left[u^2\right]} \right) \text{(dB)}. \tag{4.19}$$

The windowed HNR is:

$$\tilde{\sigma}_N(p) = 10 \log_{10} \left(\frac{\langle \tilde{v}^2(n)\, x^2(n-p) \rangle}{\langle \tilde{u}^2(n)\, x^2(n-p) \rangle} \right) \text{(dB)}. \tag{4.20}$$

4.3.2 Fundamental frequency extraction

This section considers several techniques for estimating the fundamental frequency of a speech signal, particularly with regard to subsequent analysis that we wish to perform on the signal. A survey of the most popular methods of pitch extraction would include harmonic selection (Parsons 1976), peak-picking (Howard and Fourcin 1983) and zero-crossing (Awan and Frenkel 1994), cepstral estimation (Noll 1967), inverse filtering (Markel 1972; Rothenberg 1973), maximum likelihood methods (Wise et al. 1976; Paul 1979), analysis-synthesis error minimisation (Griffin and Lim 1985), and high-resolution Fourier estimation (Brown and Puckette 1993). Only a selection was employed in this study, whose principles are explained below.

Researchers have attempted to capture the glottal motion by a variety of means: photoglottography (PGG), electroglottography (EGG, Scott and Gerber 1972; Lim et al. 1978; Rothenberg 1983; Hirose and Niimi 1987; Rothenberg 1992), and even radar (Holzrichter et al. 1998). Pitch tracking from some of these signals is easier than from the far-field acoustic pressure signal; EGG in particular is sometimes recorded specifically for this purpose, since the equipment is easy to use. This is not so true of PGG, although it still simplifies pitch extraction if the signal is available. Nonetheless, there are methods for tracking f_0 using just the recorded sound pressure signal, some of which are robust and give good precision.

Zero crossing and peak picking

Zero crossing and peak picking are ways of selecting specific points in the speech signal that correspond to some regular event, such as a glottal pulse, which can then be marked. The time difference between adjacent markers yields the pitch period. Zero crossing identifies the time instants when the speech signal changes in sign and is usually restricted to one direction, for instance when a positive sound pressure goes negative. Peak picking identifies local extrema in the signal, either maxima or minima. These two methods are in fact different forms of the same problem, since one is merely the derivative of the other; when x is a maximum, $dx/dt = 0$.

Normally a speech signal will have many peaks and zero crossings within each glottal cycle and so, for automatic pitch extraction, some additional processing is required, like low-pass

filtering or clipping (Howard and Fourcin 1983; Childers 2000). During consistent voicing that is uncluttered by noise interference, the manual selection of local maxima can provide accurate pitch marks corresponding to the same stage in the glottal cycle for consecutive pitch periods. Otherwise, information from other sources has to be incorporated to give reliable results, such as approximate pitch marks or rules to govern the amplitude and spacing of likely peaks. These techniques are therefore best used under supervision as an analysis aid, or in conjunction with more robust methods.

Autocorrelation

The equations of the autocorrelation function are defined as follows, for continuous time:

$$R_{xx}(\tau) \quad = \quad \mathrm{E}\left[x(t)\, x(t+\tau)\right] \,. \tag{4.21}$$

and for discrete time:

$$\begin{aligned} R_{xx}(m) &= \mathrm{E}\left[x_n\, x_{n+m}\right] && (4.22)\\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x(n)\, x(n+m)\,. && (4.23) \end{aligned}$$

Being a symmetrical function, the autocorrelation loses the sense of causality, which is equivalent to the absence of phase information in its frequency domain counterpart, the power spectral density:

$$\begin{aligned} S_{xx}(k) &= \frac{1}{N} \sum_{n=0}^{N-1} R_{xx}(n) \exp -j\frac{2\pi nk}{N} && (4.24)\\ &= |S(k)|^2 \,, && (4.25) \end{aligned}$$

where $S(k)$ is the Fourier spectrum of x.

The pitch period T_0 corresponds to the time lag τ at which the peak in $R_{xx}(\tau)$ occurs for the longest repeating part. In cases of diplophony or severe shimmer, and at voicing transitions (onset and offset), a subharmonic or higher harmonic peak may have an amplitude exceeding that of the fundamental, which can lead to octave errors in the estimated T_0, and hence in f_0. This unfortunate property requires that the autocorrelation method also be supervised. Moreover, note that windowing the speech prior to calculating the autocorrelation can bias the results, as pointed out by White (1997), among others.

The signals in Figure 4.1 are shown with their autocorrelation function (third from top). For the vowel [ɑ], we see a significant effect of the first formant F1 as ripple in the autocorrelation. Note that, in this case, a simple peak-picking algorithm would have correctly identified the pitch period ($T_0 \approx 8\,\mathrm{ms}$). For [z], the amount of (mostly high-frequency) noise in the signal would have produced an error in T_0 (7.3 ms vs. 7.6 ms), which could have been reduced by passing it first through a low-pass filter. However, the effect of windowing would continue to

bias the result. Pitch extraction using the autocorrelation function tends to work better for high-f_0 speech.

Cepstrum

The *cepstrum* is an alternative time representation to the autocorrelation where the magnitude spectrum undergoes a logarithmic operation before being inverse Fourier transformed, thus (Deller et al. 1993):

$$C_x(\tau) = \int_{-\infty}^{\infty} \ln|S_{xx}(\omega)| \exp j\omega t \, dt \tag{4.26}$$

$$C_x(m) = \sum_{n=0}^{N-1} \ln|S_{xx}(k)| \exp\left(j\frac{2\pi nk}{N}\right). \tag{4.27}$$

It was inspired by the desire to separate the source and filter characteristics that are effectively convolved in the time domain, taking advantage of their very different spectral forms: the VTTF is predominantly smooth in the frequency domain, whereas the glottal source is quasi-periodic and comb-like. Their cepstra are concentrated, respectively, at low time and near the pitch period, which is referred to as the first *rahmonic*. Perturbations of the voicing source produce smearing at the first rahmonic and components at multiples, the higher rahmonics. Hence, in conjunction with a restricted peak-peaking algorithm, the appropriate rahmonic can be selected, returning the estimate of f_0.

The real cepstra in Figure 4.1 (bottom) give a hint as to the sort of problem that might be encountered with this technique: the vowel cepstrum (left) exhibits a clear spike at c. 8 ms, whereas the cepstrum of the fricative has no obvious pitch peak. Both cepstra contain significant components in the low-time region (< 2.5 ms) that are responsible for the smooth spectral features, such as the formants, although the two are quite different in detail.

Subharmonic summation

The method of subharmonic summation proposed by Hermes (1988) eliminates octave artefacts by combining the harmonics on a logarithmic frequency scale that robustly achieves a global maximum at the true f_0. Several manipulations are applied to hone the spectrum before the summation operation: truncation, smoothing, interpolation and frequency weighting. Truncation eliminates spectral contributions that are more than two bins away from a local maximum by setting those coefficients to zero. The spectra are smoothed by a 3-point Hann filter (0.25, 0.5, 0.25), so as not to disturb the frequency of the peaks. These cleaned spectra are interpolated at 48 log-spaced points per octave by cubic splines, to give $A(s)$, where $s = \log_2 f$ is the log-frequency scale. Finally, a high-pass filter (raised arctangent, $f_c \approx 60$ Hz) gives a low-frequency auditory weighting $W(s)$ that removes unwanted background noise, wind noise

and interference. After pre-processing in this way, a series of $(H-1)$ shifted spectra is added to the unshifted spectrum, $W(s)\,A(s)$:

$$J(s) = \sum_{h=1}^{H} k^{h-1}\, W\left(s + \log_2 h\right)\, A\left(s + \log_2 h\right)\,, \tag{4.28}$$

where $H = 15$ is the number of harmonics considered and $k = 0.84$ is the compression factor. The global maximum in the summed function $J(s)$ gives the value of s, which provides an estimate of $f_0 \pm 0.7\,\%$. The whole process can be summarised in four steps:

1. calculate spectrum;
2. log-warp both axes;
3. shift and add for H harmonics;
4. find global maximum.

Model-matched methods

The worth of the variously extracted pitch estimates depends to a large extent on how they are to be analysed thereafter. Pitch-scaled pitch extraction, which will be described in detail in Section 5.3.2, is one of a number of methods based on minimisation of a modelling error. It has been used extensively in this study because the cost function is designed to give a pitch estimate that specifically optimises the performance of the subsequent analysis. The purpose of this note, however, is merely to alert the reader to the existence of such model-matched methods.

4.4 Inverse filters

Within the source-filter paradigm, it is a natural goal in the analysis of sources in speech to want to remove the acoustic effect of the filter. Such a problem involves estimation of the filter and application of its inverse to the recorded speech signal in order to predict the source signal. Ideally, it would provide an independent source signal, which would enable its behaviour to be investigated under various conditions. The efforts of researchers to try to deduce a "source waveform" have fallen into two distinct traditions: numerical and experimental. The numerical methods have concentrated on attempts to fit a mathematical model to observations of speech, which is then inverted and applied to the signal; experimental methods range from direct measurement of vocal fold vibration, to deductive techniques using volume velocity and pressure measurements at the lips. The deductive techniques which rely on a flow measurement, such as by the Rothenberg mask (Rothenberg 1973; 1981; Shadle et al. 1999), or inferences from the far-field sound pressure will not be discussed in this report.[2]

[2]The volume velocity recorded at the lips is inverse filtered to obtain an estimated glottal flow waveform that is maximally smooth, by adjustment of the frequency and bandwidth of a series of anti-resonators. The all-zero

Having a measurement of the glottal activity that is independent of the acoustic pressure provides a means to extract the filter characteristic from voiced speech signals without having to make any assumptions about the spectral characteristics of the voice source. The process of removing the source signal, that was convolved with the vocal tract's impulse response to produce the speech signal, is called *deconvolution*, which can be done using a least mean squares (LMS) approach with regularisation. Alternatively, we can use the signals' auto-spectra, S_{xx} and S_{yy}, and the cross-spectra, S_{xy} and S_{yx}, in a Wiener filter W (see Figure 4.2) to estimate the VTTF, which is defined as:

$$W = R_{xx}^{-1}\, p\,, \tag{4.29}$$

where R_{xx} is the autocorrelation matrix, which is symmetric and diagonal, and p is the cross-correlation vector formed from R_{xy}.

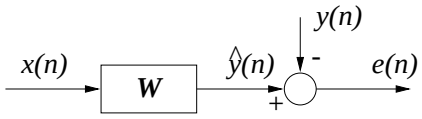


Figure 4.2: Wiener filter architecture.

4.4.1 Auto-regressive (AR) models

Linear prediction is a method of modelling a system by a weighted combination of its inputs and previous outputs. By assuming that the speech signal y is the convolution of a white (i.e., spectrally-flat) excitation signal x with an infinite impulse response (IIR) filter,

$$y_k = b_1\, y_{k-1} + b_2\, y_{k-2} + \ldots + a_0\, x_k\,, \tag{4.30}$$

linear prediction analysis makes a minimum mean squared error (MSE) estimate of the filter coefficients. Then, the excitation signal can be computed from the speech signal by applying the inverse filter. In this way, any signal that has consistent spectral colouring (over an analysis frame) can be whitened, and the filter coefficients can be used to describe the signal's spectral characteristics. The poles of the linear prediction coefficients (LPC) can be calculated by taking the roots of its polynomial (Eq. 4.30), which specify the resonances of the system, their centre-frequency and their bandwidth, and can be related to the formants of the vocal tract.

The derived excitation signal is generally noisy in appearance, but during steady voicing a regular train of spikes appears in time with the pulsing of the glottal source. This signal can be used to test for voicing or as a modified representation of the source signal, from which other features, like aspects of voice quality, may be derived. LPC can be used as a pre-processing stage

inverse is therefore equivalent to an auto-regressive model, but with the objective of smoothness rather than spectral flatness.

for efficient coding of speech signals, e.g., for mobile telephony, since the spectral artifacts of quantisation are minimal for a white excitation, whose representation can be further simplified in the absence of voicing. This property of LPC makes it a valuable pre-processor for many other signal processing techniques, including speech analysis algorithms.

4.4.2 Auto-regressive moving-average (ARMA) models

The model based on linear combinations of the current input and previous outputs can be extended to include past values of the input:

$$y_k = b_1\, y_{k-1} + b_2\, y_{k-2} + \ldots + a_0\, x_k + a_1\, x_{k-1} + a_2\, x_{k-2} + \ldots\,, \tag{4.31}$$

which is called an auto-regressive, moving-average (ARMA) model, since the weighted combination of inputs x_k is equivalent to an averaging function that slides along one point at each iteration. As well as system poles, this filter has zeros, defined by the roots of the polynomial in the coefficients a_i. Since the vocal-tract transfer function contains anti-resonances in its response, particularly for supraglottal sources, an ARMA model is more realistic than an AR model. Akin to system poles, the zeros define the centre frequencies and bandwidths of anti-resonances, augmenting the capability of the model to characterise the VTTF. However, interference from noise disguises the presence of zeros, and tends to make the inversion ill-conditioned, which can lead to spurious results. Nevertheless, regularisation can reduce the effect of these artefacts and can guarantee stable inversion, so that an excitation signal can be computed from the speech signal.

4.4.3 Electroglottography (EGG)

The experimental approach of attempting directly to measure the motion of the vocal folds shows obvious potential benefits, but each type of device has disadvantages. Photoglottography (PGG), which measures the transillumination of the glottis, depends on a sufficiently strong light source being projected onto the larynx. This is achieved by passing an endoscope through the nose and down the back of the mouth, which requires medical supervision, and the larynx can suffer from overheating from the light. Nevertheless, the intensity of the light passing through the glottis (subject to reflection off the vocal folds) can be calibrated to give a reasonable estimate of the glottal area, which can be used to predict the glottal flow (Kitzing 1983). A more practical alternative is electroglottography (EGG), which measures the electrical transconductance of the glottis using a pair (or two pairs)[3] of electrodes placed on the outside of the neck, either side of the larynx. By monitoring the current through the electrodes for a

[3] A better quality signal can be obtained using twice the number of electrodes (Rothenberg 1992), which can also discern the vertical position of the larynx, which is strongly associated with pitch gestures.

constant voltage at 2 MHz, a signal is obtained that is roughly inversely proportional to the separation of the vocal folds, which gives a good indication of the degree of contact and can be used as an approximation to the glottal area waveform (Childers et al. 1983; Rothenberg 1992).

The waveforms produced by both methods give accurate information about the timing of the glottal source (Cranen 1991). The fundamental frequency, the OQ and jitter are features that can easily be quantified, and shimmer can be estimated using the derivative of the EGG signal.

4.5 Features of plosives

The features of plosives tend to be highly transient in nature, making their analysis somewhat problematic. However, by averaging a number of repetitions of the same phoneme in similar contexts, we can eliminate some of the sources of variability and accentuate the features of interest. In this section, we examine the plosive noise at release of labial, alveolar and velar stop consonants, and the progression of sounds following the release of the labial.

4.5.1 Burst spectra

When dealing with an ensemble of tokens, differences in the speaking rate dictate that the tokens be realigned at each critical event. For the plosive release of stop consonants, this point was taken to be at the burst. Figure 4.3 shows the ensemble spectra obtained for unvoiced plosives articulated at three different places: labial /p/, alveolar /t/, and velar /k/.

Using the frequencies of the troughs in the burst spectrum, it is possible to estimate the location of the occlusion for the stop consonant. The troughs approximate to the half-wave resonances of a tube equal in length to that from glottis to source, which is assumed to be located at the obstruction. For the bilabial plosive [p], the troughs at 1.1 kHz, 2.2 kHz, 3.3 kHz, and 4.4 kHz (Fig. 4.3, top) can be explained by a source that is 16 cm from the glottis, which corresponds to the lips, for subject PJ. The patterns for [t] and [k] for which we would predict zeros at multiples of 1.3 kHz and 1.6 kHz respectively, are less obvious. One possible explanation is that the burst transient has been masked by the ensuing frication noise: for [p] the frication is co-located and weak, whereas for [t] and [k] it is stronger and likely to be located downstream of the constriction. Also, for [p], the tongue is likely to be down (especially in the context [p^hɑ]), so the simple tube model is good; for [t] and [k], the tongue forms the constriction and so this kind of model is less appropriate.

Comparing the burst ensemble spectra of unvoiced stops to the spectral envelope of their voiced counterparts reveals further differences, some of which may be explained by the effects of the anti-resonances. Stevens and Blumstein (1978) computed the spectral envelopes by pre-emphasised LPC analysis of a 26 ms window of speech, positioned over the burst. Although

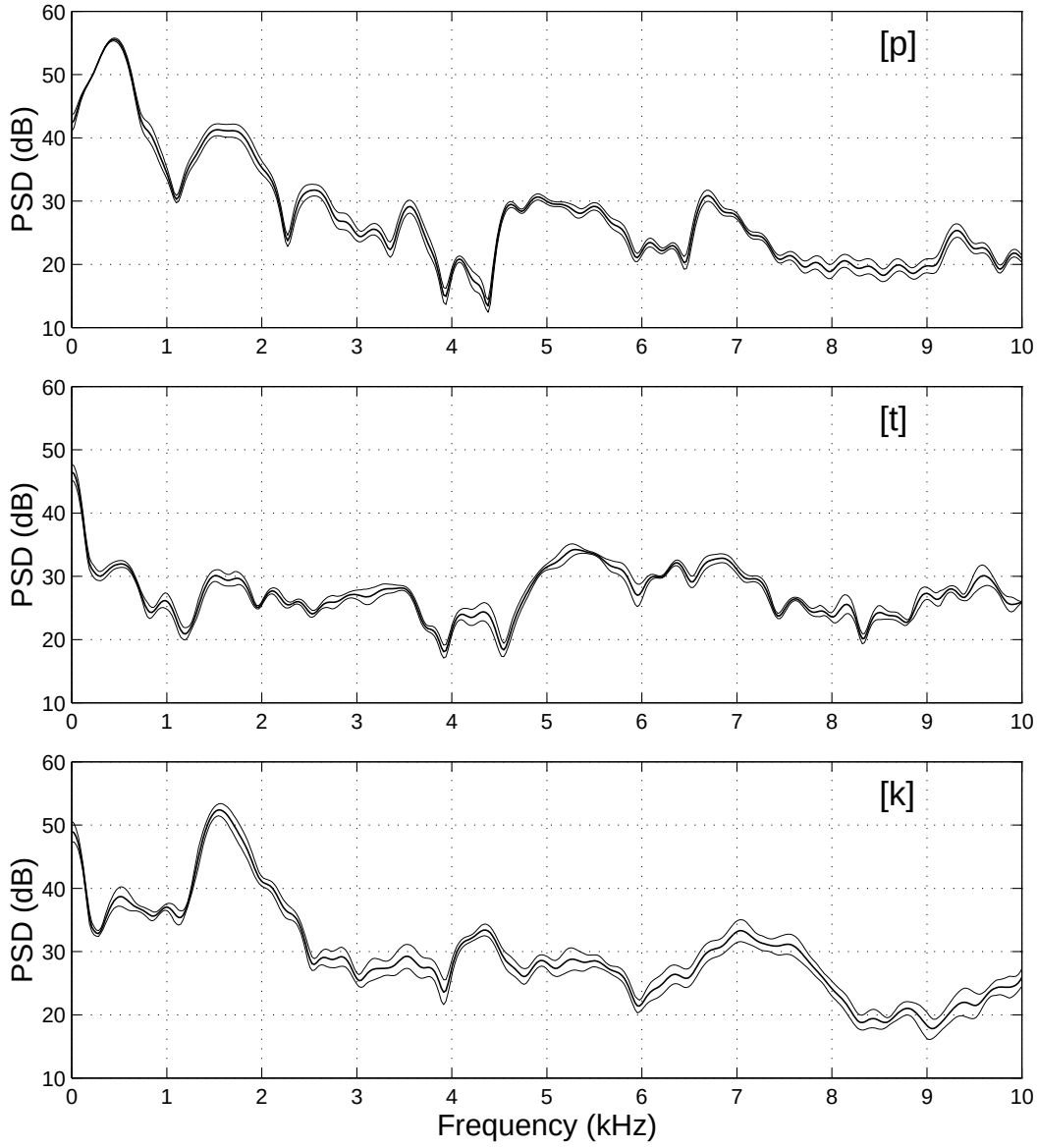


Figure 4.3: Ensemble-averaged spectra (thick line, 8 tokens, 512-point Hann window ≈ 11 ms, $\times 4$ zero-padding) of the bursts from plosive releases, with error bands (thin lines): (top) $\mathcal{C}3$-[p], (middle) $\mathcal{C}3$-[t] and (bottom) $\mathcal{C}3$-[k].

the spectral envelopes were calculated from an all-pole model, such a model is capable of characterising the voiced part of the signal well. In this respect, the slight differences in formant frequencies may be attributed to inter-speaker variation, but larger differences, such as the absence of the strong F4 peak of [d] in the [t] spectrum, are more suggestive of the influence of the source location's spectral zeros. However, there are strong similarities between the [t] ensemble spectrum and that of [s] which is articulated in the same place as [t].

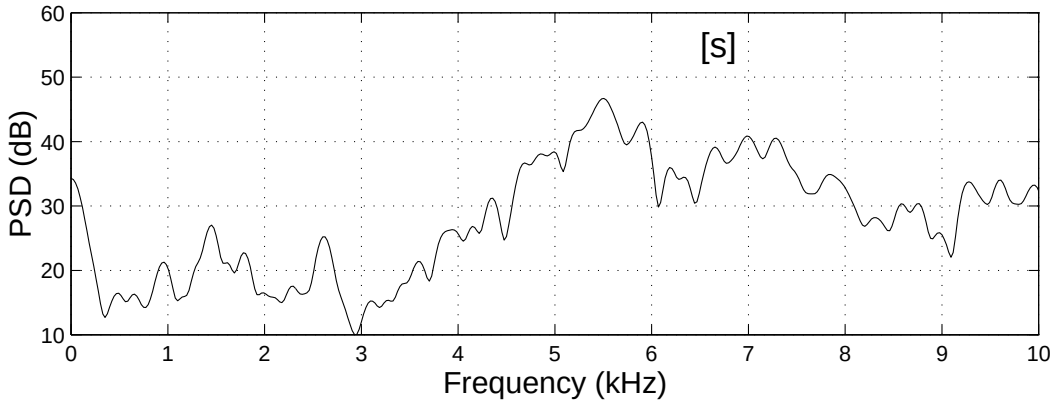


Figure 4.4: Ensemble-averaged spectrum (thick line, 8 tokens, 512-point Hann window ≈ 11 ms, $\times 4$ zero-padding) of mid-fricative /s/ in *CV*-[p$^{\text{h}}$ɑsɑ] context by PJ.

Were we to compare the [t] spectrum directly to that of [s] (as in Fig. 4.4), we would see that the main features match very well. Formants at 5.3 kHz and 6.8 kHz with broad bandwidths for [t] correspond to 5.5 kHz and 7.0 kHz in [s]. There are lesser peaks at 1.4 kHz in both, preceded by a trough at 1.2 kHz in [t], 1.1 kHz in [s]. The most striking difference overall is the spectral tilt: [t] is flatter, rising ~ 6 dB over the range 2–6 kHz; [s] rises by 16 dB over the same range. The difference is presumably a consequence of the different characteristics of the source functions.

There are also some features that appear constant across place for the unvoiced labial, alveolar and velar plosives, namely the spectral zeros at 3.9 kHz and 5.9 kHz. These may be the result of side branches, such as the pyriform sinuses, whose configuration remains unaffected by the changing position of the lips and tongue tip (Dang and Honda 1997).

4.5.2 Development

In the plosive-vowel syllable [p$^{\text{h}}$ɑ], there were several alignment markers: one at the release of the stop consonant, a second one at voice onset, and others at subsequent glottal pulses. The analysis frames between the release burst and onset were evenly spaced for each token (21 ms window, 16 tokens), and so the time offset between frames was not constant across all tokens. As a result, each averaged spectrum does not correspond to a precise time, but the average offset between frames was approximately equal to 5 ms. Figure 4.5 is a waterfall plot

that illustrates the progression in the spectra from a frame centred on the first event (release) to one on the second (voice onset) using a total of 15 frames.

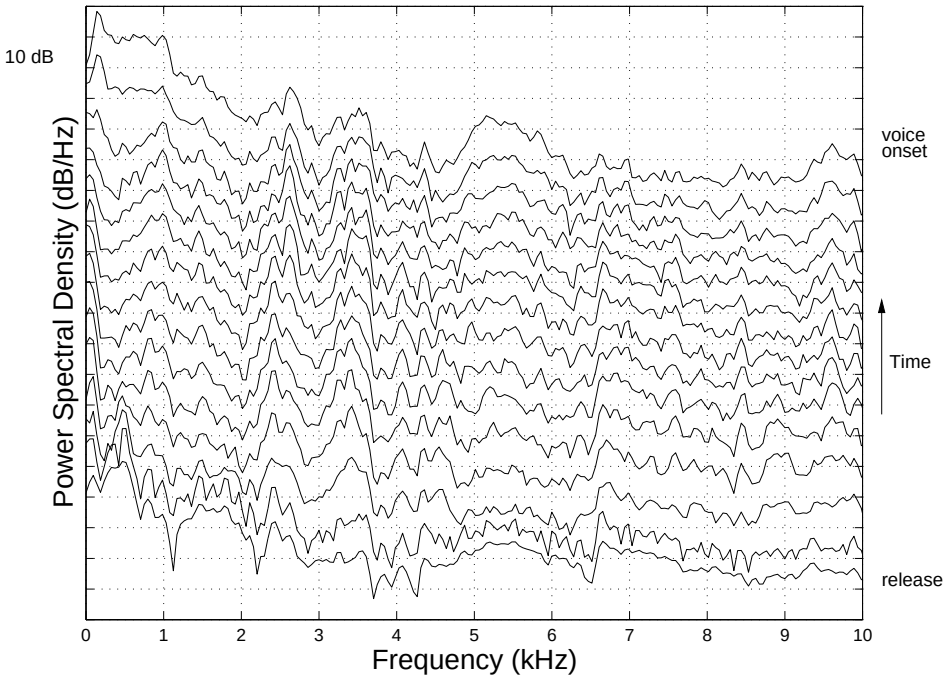


Figure 4.5: Ensemble-averaged spectra from release of *C3*-[p^h] to voice onset (16 tokens, 21 ms Hann window), in [p^hɑFɑ] context, 10 dB between tick marks and between each frame.

The deep troughs in the burst spectrum (bottom curve), which occur at the anti-resonances of the front cavity, are indicative of a source localised at the lips. Thereafter, the succeeding frames shift radically. Just after release (bottom), the formants rise in frequency (e.g., F1 from 0.5 kHz to 1.0 kHz and F3 from 2.3 kHz to 2.7 kHz), as expected due to the disappearance of lip rounding. There are also other high-frequency features which are of limited duration, e.g., the peak at 6.7 kHz during the third and fourth frames.

Band-pass filtering at the third formant (F3 $\approx$ 2.6 kHz) is often used as a means of estimating the fricative or aspirative element of a speech signal, and indeed its level does increase in the three frames after release. For this example, though, F4 $\approx$ 3.4 kHz gives a cleaner result, rising sharply two frames after release, peaking one or two frames later and gradually decaying after that. Other high-frequency peaks exhibit momentary contributions, for example at 7.6 kHz in the sixth frame, 8.9 kHz in the fourth and fifth frames, and 9.6 kHz just afterwards. The shape of the very low frequency region of the spectrum ($f < 200$ Hz) may also provide some useful clues, since it changes shape immediately after release and again at voice onset. It may be possible to distinguish particular regions as associated with aspiration rather than frication (or vice-versa), but without additional information we can only note that different frequencies are excited at different times during the progression of sounds from the release of the plosive, through frication and aspiration, to the onset of voicing.

4.6 Summary

This chapter describes the procedure and content of the speech recordings giving rise to the corpora used for analysis in the present study. Some traditional analysis techniques were introduced, and various parameters and models were defined that are used for speech analysis. It was shown how features extracted from an ensemble-averaged burst spectrum (that are normally overlooked by LPC analysis) could be used to identify the source location. A collection of synchronised plosives was combined to give the changing pattern of sound through the development of the stop consonant, from the release to voice onset. Despite the range of analysis possibilities, however, mixed-source sounds present a problem for investigating the characteristics of individual sources. For voiced sounds, we can use the predictable nature of the vocal oscillation to achieve a separation of the voiced and unvoiced components, thus enabling us to examine the contributions from different sources individually. In the next chapter, methods for decomposing the speech signal will be discussed.

Chapter 5

Decomposition of mixed-source speech: Method

5.1 Introduction

It has been noted that aspiration noise is almost always present in speech and, being generated near the glottis, it is likely to interact strongly with voicing. Being dependent on the flow velocity, frication is also influenced by vibration of the vocal folds. As we have seen, there are interwoven patterns of acoustic events at the release of a stop consonant, as there are at the initiation and termination of voicing. During voiced stops, the closure obstructs the flow of air through the glottis and tract, which dramatically affects the production of both voicing and turbulence noise. The majority of speech therefore contains simultaneous contributions from more than one type of acoustic source. By considering the relative levels of contributions from voicing and from bursts and turbulence noise, we can provide some kind of description using the harmonics-to-noise ratio (HNR). To study these phenomena properly, we would like to be able to analyse the voiced and unvoiced components of mixed-source speech separately, possibly even to distinguish between all the different contributions. For each source, we would like to describe its characteristics and to explore its properties, in particular where and how it is produced. Signal processing techniques have been developed for decomposing speech signals into quasi-periodic and aperiodic components, which can be considered to be estimates of the voiced and unvoiced parts, respectively. Ideally, the periodic or harmonic component would contain precisely the vocal-tract-filtered voicing source, and the aperiodic or anharmonic component the filtered noise sources. These signals can be used for comparison of source envelopes and the synchrony of articulatory events, and potentially for automatic identification of the type and location of source, e.g., via frequencies of peaks and troughs in the anharmonic spectrum.

This chapter describes a selection of techniques that can be enlisted to separate the voiced

(harmonic) and unvoiced (anharmonic) components during phonation, and describes in detail a technique that we have developed. The method that we propose, the pitch-scaled harmonic filter (PSHF), provides four reconstructed time series signals by decomposing the original speech signal, first, according to the signal amplitude, and then according to its power. The result is one pair of decomposed (harmonic and anharmonic) signals optimised for time-series analysis, and another pair for spectral analysis.

It is well-known that many acoustic cues are asynchronous in speech production, coming from composite articulatory trajectories whose gestures are not perfectly coordinated (sometimes referred to as 'non-linear' by phoneticians). Much attention has been given to separating and labelling acoustic cues in large corpora of speech signals. By increasing the number of concurrent signal strands from one to four, it is anticipated that the capability for asynchronous cue timing is increased considerably.

In this chapter, the performance of the PSHF algorithm is compared against other techniques using synthetic signals, and evaluated under three forms of signal perturbation: jitter (perturbed fundamental frequency, f_0), shimmer (perturbed amplitude), and additive noise with variable burst duration. The results of these tests can be employed to predict the performance on real speech. In Chapter 6, we give examples from speech recordings that were analysed to illustrate some of the decomposition technique's practical benefits. The periodic-aperiodic decomposition (PAPD), which is an alternative technique based on a cepstral HNR measure (de Krom 1993), is discussed in detail in Appendix D.

5.2 Review of decomposition methods

Further to the discussion in Chapter 1, this section explores the predominant approaches to decomposing speech signals. Then, it suggests a way that scaling the analysis window to f_0 can directly improve the results of decomposition.

5.2.1 Time domain (TD)

It is generally accepted that voicing produces a series of glottal pulses which excite the vocal tract, and that these are quasi-periodic in normal, sustained phonation. This property may be exploited to produce an estimate of the voiced part by aligning the acoustic responses to a number of pulses and averaging out other variations, such as turbulence noise produced by unvoiced sources. Unfortunately, the pulses are not perfectly timed and any variations in periodicity which are not modelled also contribute to the residual, unvoiced components. Early attempts to accommodate such variations averaged the signal only over a small number of periods, giving more weighting to the central period, and less to the extremities (Shields

1970; Lim et al. 1978). Thus, a short time-window was used to give an adaptive comb filter. However, timing variations within the windowed frame continued to cause errors, to which a number of solutions has been proposed. Frazier et al. (1976) used the duration of each pitch pulse to determine the spacing of the comb filter's teeth, and matched the length of the periods for averaging by truncation or zero-padding, as appropriate. Yumoto (1982) used a phase normalisation procedure to make all periods have the same duration. Pinson (1963) performed a least-squares alignment of successive periods, which was later reformulated into a maximum likelihood optimisation (Feder 1993), and a dynamic time warping problem (Graf and Hubing 1993). By dealing with each period in terms of its Fourier series, Murphy (1999) effectively stretched the time scale linearly to align the end points of each cycle, and by also normalising the magnitude, he addressed the issue of variations in amplitude. These approaches failed to recognize that, while the vocal fold oscillation (and hence the glottal source waveform) may exhibit such gross distortions, the vocal-tract impulse response is unlikely to vary in the same way. Nevertheless, the effects of changes in vocal-tract configuration may be largely considered as slow and hence neglected to a first approximation.

It is evident that time domain methods rely heavily on accurate pitch-timing information but, being a requirement of many analysis techniques, various solutions have been developed, which were discussed in Chapter 4. TD comb filters have the advantage of low computational complexity, since no transformation of the input signal is required. However, the errors resulting from the variability of real speech produce glitches in the output waveforms close to pitch pulses and smearing of the high frequency components. Indeed, careless processing can leave the decomposed components sounding somewhat mechanical. Yet the main drawback is that the effects of these errors are difficult to understand, particularly in frequency spectra.

5.2.2 Frequency domain (FD)

Frequency domain methods are generally more suitable when one is considering FD features as one's final representation of the signals, e.g., peaks and troughs in the spectral envelope. Instead of smearing the high frequency region of the spectrum (as with adaptive comb filtering), it can be preserved, but the consequence is that the pitch harmonics themselves become smeared by the effects of jitter and shimmer. Features of this space are more likely to belong to the vocal-tract transfer function (VTTF) than the source function, and so FD methods are particularly suited to studies of frication and aspiration, which are predominantly characterised by the turbulence noise's spectral shape.

When using the asynchronous STFT spectra to calculate the voicing model, there are bias terms that negate the optimal properties of the discrete Fourier transform (DFT) for frequency analysis. These will be discussed in more detail later, in Section 5.2.5, where we

consider a pitch-scaled approach. Otherwise, high-amplitude harmonics will tend to interfere with their neighbours and, if they are not sufficiently far apart in frequency, to distort the model. Moreover, f_0 can be biased by its own image at $-f_0$. These effects can be reduced by prudent choice of window function and frame length, but if one is prepared to assume a perfectly periodic model, these factors can be explicitly removed asynchronously using a suitable matrix inverse (Silva and Almeida 1990; White 1997). As mentioned in the Introduction (Ch. 1), pitch-scaled processing avoids these problems.

Among the methods taken from speech enhancement, Hardwick, Yoo and Lim (1993; Yoo and Lim 1995) used decomposition to enable different approaches to the enhancement of the voiced and unvoiced components (their dual excitation model). A classic enhancement task is that of separating the speech of two simultaneous talkers, the cocktail party problem. One approach is to have two voicing models and extract the relevant coefficients from the signal for each, assuming the remainder to be noise. For FD models, this implies the identification and tracking of the two f_0 trajectories, which specify the corresponding harmonic spacings (Parsons 1976; Silva and Almeida 1990; Damper et al. 1995). Of course, by treating the unvoiced components as interference, the models provide no mechanism for separating those sounds, although good separation of voiced components will aid a human listener to do so by augmenting other cues.

Stylianou developed a suite of techniques for speech modification that start from the classical premises of FD models (Stylianou 1995; Stylianou et al. 1995). He extended the modelling of harmonics to account for linear f_0 trajectories over the course of an analysis frame. Rather than merely allowing smooth adaptation between frames, pitch changes within a frame were explicitly modelled, thus eliminating the stationary condition. In his deterministic-stochastic model, even the harmonic relation of tones is relaxed so that changes need not be the same for all f_0 multiples (although the mid-frame frequencies were harmonically related). He also experimented with alternative models of the unvoiced components, based on 'noisy' and 'stochastic' assumptions. In Laroche, Stylianou, and Moulines (1993), linear f_0 variation was included within a frame, but in their demonstration (pitch-synchronous, two-period window) the data were over-parameterised, resulting in 3 kHz low- and high-pass filtered representations of the voiced and unvoiced components, respectively.

5.2.3 Correlation methods

A couple of correlation-based methods have been published. Michaelis et al. (1995) divided the speech signal into a number of frequency bands and determined whether each band was voiced or unvoiced, rather like Griffin and Lim (1988), but on the basis of the correlation between the bands. Qi and Hillman (1997) extracted the short-term and long-term correlations from

the signal to remove the effect of the VTTF and the periodicity of voicing, respectively. Thus, the residual acted as an estimate of the unvoiced excitation, which could be used with the short-term correlation to give the unvoiced signal. Similarly, the estimate of the voiced part could be rebuilt from both of the extracted correlations.

5.2.4 Cepstral methods

A collection of cepstral methods has emerged in recent years, e.g., Yegnanarayana et al. (1998), and Qi et al. (1999). Most of these methods use a cepstral filter that passes narrow time-bands of the cepstrum centred on the pitch rahmonics (de Krom 1993). Like all methods, it has its limitations, such as being unable to support gross differences in spectral shape (such as have been observed, Jackson and Shadle 2000c). There is no compensation for the unvoiced part in the rahmonics, and so neither is this approach effective at low HNRs, as it stands. One variant (Yegnanarayana et al. 1998) is discussed in detail in Appendix D.

In their adaptations of the de Krom technique, Qi and Hillman (1997) removed the cumbersome baseline shifting operation, but their results were disappointing. The version by Yegnanarayana et al. (1998) bypassed this issue, by aiming to decompose the signals directly, rather than making a comparison of their amplitude levels. They suggest (but do not implement) using the cepstra of both harmonic and anharmonic components to decide the allocation of frequency bins to each component. Indeed, they could go further, and use the initial bin-wise ratio of the two (a bin-wise HNR estimate) to divide the bin's power, in a sense determining the most likely component spectra, given the observed spectrum. With the limited information available any probabilistic framework of this type would be quite rudimentary, but is certain to yield better estimates of the spectra, given the right assumptions.

5.2.5 A pitch-scaled approach

We use the term *pitch-scaled* to refer to an analysis frame that contains a small integer multiple of pitch periods. It implies, for a constant sampling rate, that the number of sample points in the frame will be inversely proportional to the fundamental frequency. This property complicates the windowing and re-splicing processes, but it brings substantial benefits too.

The main benefit, which we exploit, is that the harmonics of f_0 will be aligned with certain bins of the DFT (assuming we know the value of f_0). For example, if our analysis frame contains b pitch periods, then the frequency of every bth Fourier coefficient will correspond to a harmonic of f_0. When the frequency in question is not exactly aligned with one of the discrete frequency bins, leakage and spectral smearing take place. We can formalise these arguments by considering some idealised signals.

For a single infinite sinusoid of frequency f_1 in Gaussian white noise (GWN), the highest

peak in the DFT spectrum provides the least-squares estimate (minimum mean-squared error) of the magnitude, frequency and phase of the sinusoid, given enough samples are taken at a high enough rate (Rife and Boorstyn 1974; Priestley 1981). Moreover, that estimate coincides with the maximum likelihood estimate, since the peak of the Gaussian distribution occurs at the mean value of the noise (i.e., at zero, Bretthorst 1988). However, a bound on the number of samples in the frame limits the frequency resolution of the DFT, and if f_1 is not coincident with one of the bins, the spectrum displays smearing and leakage. Also, if f_1 is of the same order as the frequency resolution, the negative-frequency image centred at $-f_1$ will bias the estimates (Silva and Almeida 1990). Similarly, if the sampling rate f_s is too low, so that the Nyquist of folding frequency is near f_1, there will be similar bias effects from aliasing. In contrast, if the analysis frame is chosen to have several whole cycles (with sufficiently high f_s), so that f_1 lies on a DFT bin, the bias terms from interference of f_1 with $-f_1$ and the spectral leakage will disappear; the remaining error is unbiased Gaussian noise whose variance is proportional to that of the additive noise.

When there is more than one sinusoid present in GWN, the situation becomes more involved. To maintain optimal (maximum likelihood) estimation of the deterministic components they must be sufficiently separated in frequency with respect to the frequency resolution, as well as each meeting the earlier constraints; otherwise, biases are introduced from cross- and auto-interference terms, respectively (Rife and Boorstyn 1976; Bretthorst 1988). Again, we can avoid these terms provided the frame is scaled to the frequency of both sinusoids, which therefore requires that they be harmonically related.

However, speech signals, although predominantly harmonic, are not composed of pure sinusoids of infinite duration. Vibration of the vocal folds tends to generate sound pressure signals that are approximately periodic, but whose amplitude and fundamental frequency fluctuate during voicing and change dramatically at voice onset/offset. To accommodate such non-stationarity, we have elected to use a Hann window, which still yields unbiased estimates, although it increases the variance of the error by 50 % (Jenkins and Watts 1968; Rife and Boorstyn 1976). This step greatly enhances the technique's robustness to minor pertubations in periodicity.

5.3 Pitch-scaled harmonic filter (PSHF)

Our approach aims to separate the principal components of the speech signal, those voiced and unvoiced, making use of our knowledge of the acoustic mechanisms of sound production. By decomposing the signals, we can better study the properties of the constituents which, in the case of voiced fricatives, for instance, are voicing and frication noise. In particular, the pitch-

scaled harmonic filter was designed to separate the harmonic and anharmonic components of speech signals. It is assumed that these components will be representative of the vocal-tract filtered voice source and noise source(s), respectively. The original speech signal $s(n)$ is decomposed primarily into the harmonic (voiced) and anharmonic (unvoiced) components, $\hat{v}(n)$ and $\hat{u}(n)$ respectively. Further harmonic and anharmonic estimates, $\tilde{v}(n)$ and $\tilde{u}(n)$, are computed based on interpolation of the anharmonic spectrum, which improves the spectral composition of the signals when considering features over a longer time-frame.

This method is especially suited to acoustic analysis of sustained sounds with regular voicing, because of the underlying harmonic model of the voiced part. Other than the choice of the number of pitch periods (typical of adaptive filtering techniques), the PSHF has no arbitrary parameters requiring heuristic adjustment, such as cut-off frequency (Laroche et al. 1993) and number of cepstral coefficients (Qi and Hillman 1997; Yegnanarayana et al. 1998), and does not suffer the bias, harmonic-interference and variable performance problems of asynchronous harmonic techniques (Serra and Smith 1990; Silva and Almeida 1990; Hardwick et al. 1993; Laroche et al. 1993; Qi and Hillman 1997; Yegnanarayana et al. 1998). The ability of the technique to provide its output as time-series signals enables subsequent analyses of the components to be performed independently. Thus, the outputs can be examined using traditional analysis techniques, but the decomposition exposes opportunities to develop new methods of analysis. An example of such an analysis procedure that exploits having two simultaneous components is demonstrated in Chapter 7.

5.3.1 Origins

Our decomposition technique is based on a measure of harmonics-to-noise ratio derived by Muta et al. (1988). In the process of calculating the HNR from a short section of speech $s(n)$, they used the spectral properties of an analysis frame scaled to the pitch period for distinguishing parts of the spectrum containing harmonic energy from those without. To do this, they applied a window function w of length $N(p)$ to $s(n)$, centred at time p, to form

$$s_w(n) = w(n)\, s(n + p - N/2)\,. \tag{5.1}$$

They computed the spectrum $S_w(k)$ by DFT (where the subscript w denotes windowing) using a value of $N = bT_0$ that was a whole number b of pitch periods of length T_0 (in samples):

$$S_w(k,p) = \sum_{n=0}^{N-1} s_w(n) \exp\left(-j\frac{2\pi nk}{N}\right), \tag{5.2}$$

which concentrated the periodic part of s_w into the set of harmonic bins B, where B contains every bth coefficient: $\{b, 2b, 3b, \ldots, b(N-1)\}$. Choosing a four pitch-period Hann window,

$$w(n) = 0.5\,(1 - \cos 2\pi n/N) \text{ for } n \in \{0, 1, \ldots, (N-1)\}, \tag{5.3}$$

the harmonics are translated to bins $\{4, 8, 12, \ldots\}$. They chose $b = 4$ because four is the smallest number that leaves bins free of spectral leakage from the periodic component (those half-way between the harmonics: $\{2, 6, 10, \ldots\}$). In fact, once the pitch period T_0 has been determined, any whole number of them $b\,T_0$ can be used to compute the spectrum, and (since our technique does not directly use the spectral amplitudes of the inter-harmonic bins) the harmonics can always be extracted to generate the voiced estimate. Thus, b can potentially be any positive integer, although we have not tested any alternatives. There is inevitably a trade-off between time and frequency resolution in the choice of b which, among other things, balances the noise rejection performance against the tolerance to jitter and shimmer. However, their value of $b = 4$, which has a time-scale comparable to other adaptive techniques, e.g., Frazier et al. (1976), offers a reasonable compromise for speech signals between adaptability and ideal PSHF performance, that yields favourable decompositions. Thus, for an adult male speaker with pitch period of 7.5 ms, a window of 30 ms duration would be used.

Muta et al. (1988) used different methods for estimating the harmonics and the noise from the short-time spectra. The harmonic power was the power spectral density (PSD) integrated over the harmonic bins; the noise power in each group of four bins was taken as the minimum PSD (usually the one half-way between the harmonics), and integrated over all the bins.

Using their value of $b = 4$, we have extended the process to yield a full decomposition into harmonic (estimate of voiced) and anharmonic (estimate of unvoiced) complex spectra, which can be converted back into time series $\hat{v}$ and $\hat{u}$ respectively, as explained below. We also propose an interpolation step for improving power-spectral estimation, which produces $\tilde{v}$ and $\tilde{u}$ (Jackson and Shadle 1998, 1999b, 2000c). The signals can later be analysed using any standard technique: $\hat{v}$ and $\hat{u}$ for TD analysis, $\tilde{v}$ and $\tilde{u}$ for FD analysis. For time-frequency analysis, we define a threshold of half the mean PSHF window length, $\langle N \rangle/2$ or two pitch periods, which is the point at which the harmonics begin to be resolved. Thus, $\hat{v}$ and $\hat{u}$ would be used for wide-band spectrograms, and $\tilde{v}$ and $\tilde{u}$ for narrow-band. The remainder of this section describes the Muta et al. (1988) pitch estimator, the segmentation of speech signals into frames and the PSHF algorithm.

5.3.2 Pitch estimation

The PSHF relies on the window length N being scaled to the time-varying pitch period T_0: $N(p) = bT_0(p)$. The pitch-tracking algorithm estimates the period by sharpening the spectrum at the first H harmonics, $h \in \{1, 2, \ldots, H\}$, as in Muta et al. (1988). Their sharpness was described in terms of the higher and lower spectral spread, S_h^+ and S_h^- respectively, which are

defined for a given window at each harmonic, $h \in \{1, 2, \cdots, H\}$ as:

$$S_h^+(N,p) = |S_w(bh+1)|^2 - \frac{|S_w(bh)|^2}{|W(h\,\Delta f_0)|^2} \left| W\left(h\,\Delta f_0 - \frac{1}{N}\right)\right|^2 \tag{5.4}$$

$$S_h^-(N,p) = |S_w(bh-1)|^2 - \frac{|S_w(bh)|^2}{|W(h\,\Delta f_0)|^2} \left| W\left(h\,\Delta f_0 + \frac{1}{N}\right)\right|^2 , \tag{5.5}$$

where $\Delta f_0 = 1/\Delta T_0 = b f_s/\Delta N$,

$$W(k) = \frac{N}{2}\left(\operatorname{sinc}\pi kN + \frac{1}{2}\left[\operatorname{sinc}\pi(kN-1) + \operatorname{sinc}\pi(kN+1)\right]\right)\exp -j\pi\Delta f_0 N\,,$$

and $\operatorname{sinc} x = \sin(x)/x$. Thus, the ideal spectral smearing of each measured harmonic is computed over the adjacent higher and lower bins $k = bh \pm 1$, as is a product of windowing, and the values are compared to the measured values in those bins. The optimum pitch estimate $N(p)$ is obtained by minimising the difference between the ideal and measured smearing in a minimum mean-squared error sense, according to the cost function at time p:

$$J(N,p) = \sum_{h=1}^{H}\left(S_h^+(N,p)^2 + S_h^-(N,p)^2\right) . \tag{5.6}$$

See Muta et al. (1988) for further details. The optimisation is perfectly matched to the PSHF because, using the same window, it maximises the concentration of signal energy into the harmonic bins.

For each section of voiced speech, the initial estimate of $N(p)$ was set manually. For larger data sets, standard methods could easily be implemented for automatic initialisation, e.g., Noll (1967), Hess (1983) and Hermes (1988). The pitch tracker operated as follows: window speech signal (N-point, Hann); evaluate cost function $J(N,p)$ near current estimate; update the current estimate $N(p)$ to N_{opt} (value at minimum cost); increment time p and repeat.

Apart from the choice of window function, the amplitude of the input speech signal obviously affects the magnitude of the cost function, yet the amount of spectral spreading is only meaningful in relation to the amplitude of the spectral peaks at the harmonics. Since our pitch-tracking results were heavily supervised and each track checked manually, we were not concerned that such gross effects could affect the pitch estimates. Nevertheless, normalisation of the cost function to the PSD of the harmonics, or to the total signal power, would provide a cost that was a generic indication of the quality of the pitch estimate, which could itself perhaps be used as a measure of voice quality and converted to an HNR.

5.3.3 Windowing and re-splicing

Windowing is essential for processing finite frames or sections of data, but for this piecewise stationary model it also allows the PSHF to adapt in line with the many kinds of variation in the speech production system: amplitude, fundamental frequency, formant frequencies, voice onset/offset and other transients. After decomposition, the output signals can be recombined

by overlapping and adding. Since the decomposition algorithm is only of relevance during voicing, a practical processing unit is a single period of phonation, i.e., from voice onset to the next offset.

The quasi-periodic signal that results from variations in the pitch of real speech presents two problems for time-scaled segmentation:

- adjustment of the window size and/or shape to match the changing pitch period, and
- overlapping of the windows to ensure shift-invariant unity gain through the segmentation/reconstruction process.

Two methods were proposed. The first used asymmetric cosine windows:

$$w(n) = \begin{cases} \frac{1}{2}\left(1 - \cos\frac{2\pi}{N_1}n\right) & \text{for } n \in \{0, 1, \cdots, (\frac{N_1}{2} - 1)\} \\ \frac{1}{2}\left(1 + \cos\frac{2\pi}{N_2}\left(n - \frac{N_1}{2}\right)\right) & \text{for } n \in \{\frac{N_1}{2}, \frac{N_1}{2} + 1, \cdots, \left(\frac{N_1+N_2}{2} - 1\right)\} \end{cases} \tag{5.7}$$

with overlapping sections that were matched for unity gain throughout the period of phonation. The second segmentation method keeps the Hann window symmetric and true to the local pitch-period estimate, but normalises the final envelope of the accumulated signal using the sum of the window contributions. Thus, a smooth reconstruction can be obtained with a reasonable amount of overlap (i.e., greater than 50 %), indeed from the point of view of the signal power, at least 75 % overlap is desirable. An aggregate is built up from successive summations of the windows, amounting to an effective weighting of each sample in the speech record. The outputs were then normalised to unity gain after processing. For simplicity, the centre positions p_i of the frames i were spaced at a constant interval: $\alpha = p_i - p_{i-1}$. However, since the window size was not generally constant, neither was the signal weighting; lower fundamental frequency regions, having longer windows $w_i(n)$, accrued more weighting than higher f_0 regions. Therefore, to normalise the output signals, i.e., the reconstructed harmonic and anharmonic components, they were multiplied by $W(p)$, the reciprocal of the sum of the window weightings of all windows w_i centred at p_i:

$$W\,(p) = \frac{1}{\sum_i \{w_i\,(p - p_i + N(p_i)/2)\}}\ , \tag{5.8}$$

for all frames i (not necessarily contiguous) that included the point p.[1] A cosine ramp was applied to each end of the normalisation factor $W(n)$ to fade out sections of voicing at onset and offset.

Preliminary comparison of the two methods did not reveal any significant discrepancies, but there were concerns about the effect on the spectrum of the asymmetric windows, the

[1] Assuming that f_0 varies gradually over the interval α, a further alternative would be to normalise the area under each frame's window prior to the decomposition to give an even point-wise weighting, as in Lim et al. (1978).

reliability of their specification and the amount of overlap. So, the latter method was adopted, with a deliberately cautious policy of high overlap.

5.3.4 Algorithm

Harmonic filter

Let us consider how the PSHF algorithm performs the decomposition in the FD for a single frame, centred at time p. (Note: all functions within the algorithm are adaptive and depend on p, but for clarity, we omit the argument p hereafter.) After applying the pitch-scaled Hann window to the speech signal to get $s_w(n)$, the PSHF algorithm computes $S_w(k)$ by DFT, as depicted in the flow diagram in Figure 5.1. The harmonic filter takes the pitch harmonics from

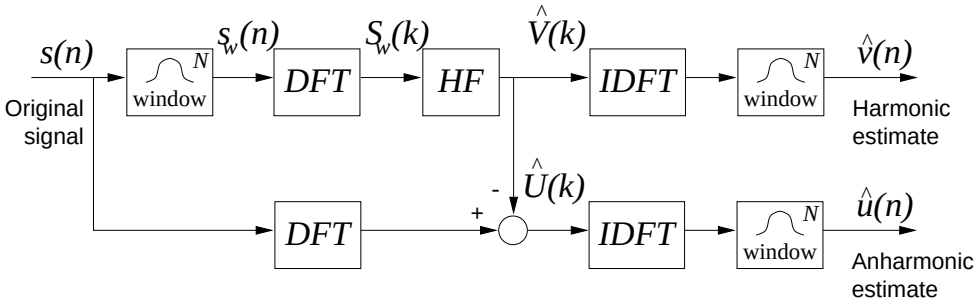


Figure 5.1: The basic pitch-scaled harmonic filter (PSHF), which comprises windowing, discrete Fourier transform (DFT), harmonic filter (HF) and inverse DFT (IDFT) operations.

S_w and doubles the coefficients to form the harmonic spectrum $\hat{V}(k)$, which compensates for the mean window amplitude of 0.5:

$$\hat{V}(k) = \begin{cases} 2S_w(k) & \text{for } k \in B \\ 0 & \text{otherwise}, \end{cases} \tag{5.9}$$

where $B = \{4, 8, \ldots, 4(N-1)\}$. This spectrum, when returned to the time domain by inverse DFT (IDFT), produces a signal that is periodic with no envelope shaping, so these four pitch periods are windowed to yield the harmonic signal estimate:

$$\hat{v}_w(n) = \frac{w(n)}{N} \sum_{k=0}^{N-1} \hat{V}(k) \exp\left(j\frac{2\pi nk}{N}\right). \tag{5.10}$$

The anharmonic signal estimate is the difference between the input signal and this harmonic estimate: $\hat{u}_w(n) = s_w(n) - \hat{v}_w(n)$. Alternatively, in the frequency domain, we can subtract $\hat{V}$ from the unwindowed spectrum:

$$\hat{U}(k) = \begin{cases} S(k) - 2S_w(k) & \text{for } k \in B \\ S(k) & \text{otherwise}, \end{cases} \tag{5.11}$$

and then the anharmonic component $\hat{u}_w$ comes from applying the IDFT and window, as before. As a result, any errors in the harmonic estimate caused by the decomposition algorithm are (wrongly) attributed to the anharmonic signal.

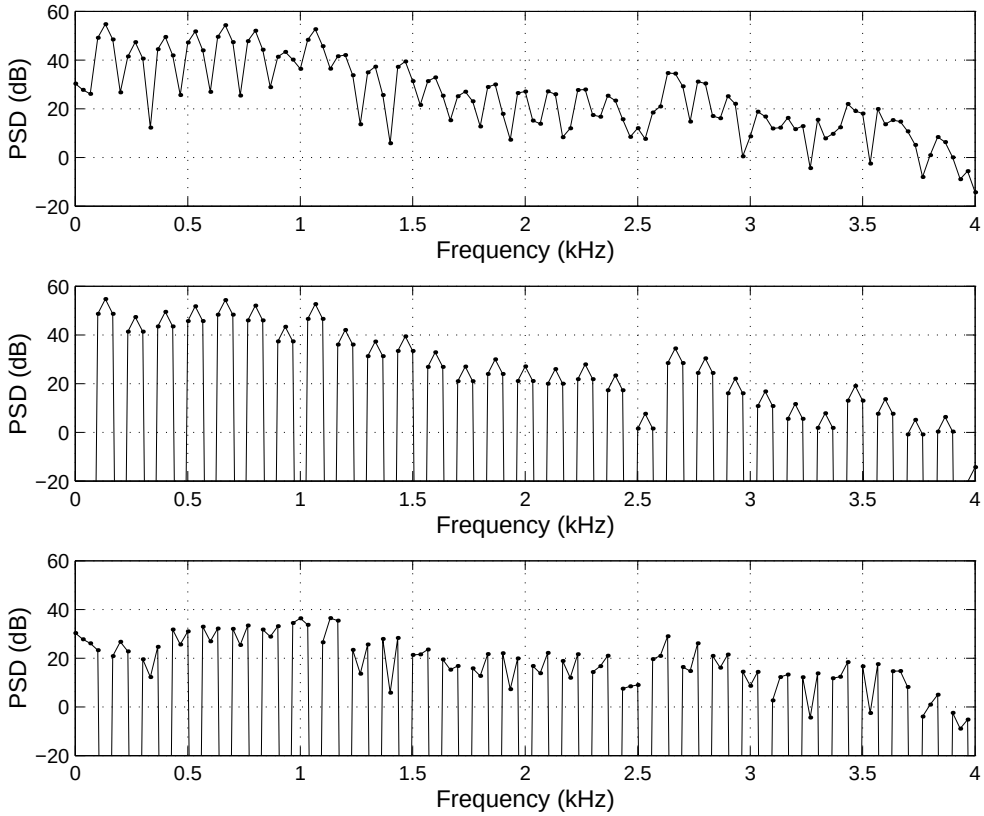


Figure 5.2: Spectra of (top) windowed speech signal $S_w(k)$, (middle) the harmonic estimate $\hat{V}_w(k)$, and (bottom) the anharmonic estimate $\hat{U}_w(k)$.

Figure 5.2 illustrates the operation of the harmonic filter by showing (a) the original spectrum $S_w(k)$, (b) the spectrum of the harmonic estimate $\hat{V}_w(k)$, and (c) the anharmonic spectrum $\hat{U}_w(k)$. The original spectrum was calculated using a mid-vowel recording of [ɑ] by an adult male (from example #1 by PJ used in Chapter 6). The time series were decimated for clarity. The essence of this technique is that, by scaling the window size to exactly four pitch periods, $N = 4T_0$, the voiced (quasi-periodic) part is concentrated into every fourth bin of the spectrum. The pitch estimation process finds the value of T_0 that optimises that concentration. Thus, a harmonic comb filter, which passes these harmonic bins, results in an optimal periodic estimate of the voiced component, of length $4T_0$. Doubling and re-applying the window matches the estimate's envelope to that of the input signal $s_w(n)$. The spectral consequences can be seen in Fig. 5.2 (middle), which shows how, for each harmonic, the Fourier coefficient maintains approximately the same value as that of the original spectrum (Fig. 5.2, top), but has spread to the adjacent bins (at $-6\,\mathrm{dB}$). The residue is the anharmonic component, whose spectrum (Fig. 5.2, bottom) accordingly contains gaps at the harmonics.

Power interpolation

The spectrum of the anharmonic signal estimate $\hat{U}_w(k)$ contains gaps at the harmonics, where the coefficients are of zero amplitude, since $\hat{U}_w(k) = S_w(k) - (2S_w(k))/2 = 0$ for $k \in B$. However, subsequent analysis often involves computing power spectra or spectrograms, which depend on the squared magnitude of the Fourier coefficients, and the gaps therefore give strongly biased under-estimates. We can improve the power estimates by filling $\hat{U}_w$ in at the harmonics, as illustrated in Figure 5.3.

If we assume that the anharmonic component is the result of a stochastic process with a smoothly varying frequency response, we would expect the power in any frequency bin to be similar to its adjacent bins. Therefore, we calculate $L(k)$, a frequency-local estimate of $|U_w|$ at the harmonics, by power interpolation (PI) of the values of the anharmonic spectrum in the adjacent bins, $\hat{U}_w(k \pm 1)$:

$$L(k) = \sqrt{\frac{\left|\hat{U}_w(k-1)\right|^2 + \left|\hat{U}_w(k+1)\right|^2}{2}} \quad \text{for } k \in B\,. \tag{5.12}$$

The RMS amplitude $L(k)$ is compared with the harmonic spectrum $\hat{V}_w(k) = S_w(k)$ for $k \in B$, to determine the real factor $\lambda(k)$, which is the proportion of the coefficient to be allocated to the revised anharmonic estimate $\tilde{U}(k)$, for each harmonic:

$$\lambda(k) = \frac{L(k)}{\sqrt{|S_w(k)|^2 + L(k)^2}}\,. \tag{5.13}$$

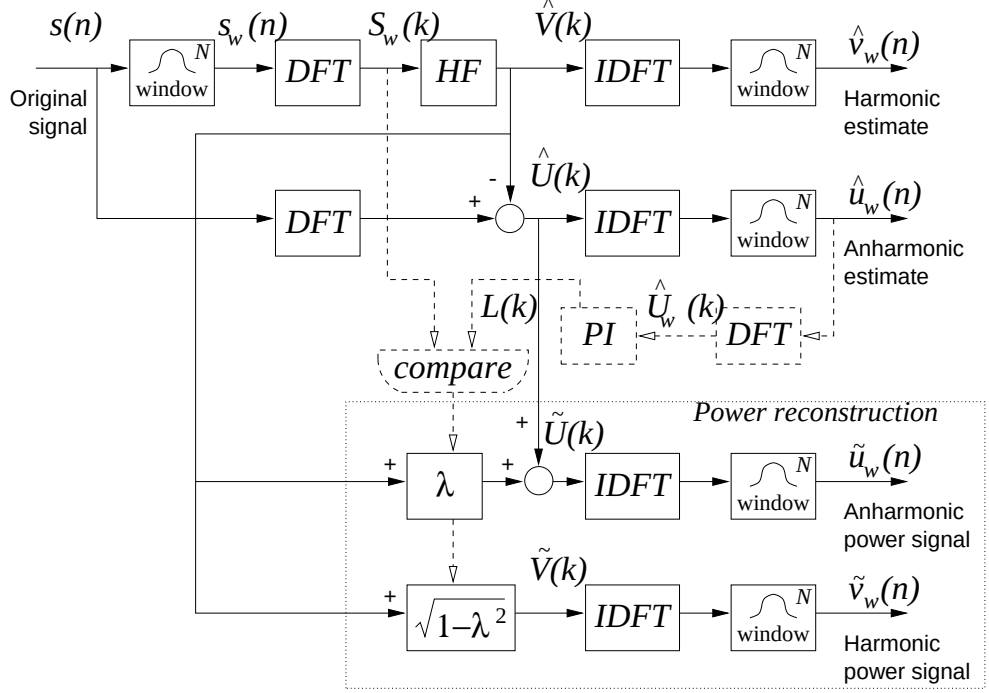


Figure 5.3: The complete pitch-scaled harmonic filter algorithm. The top half provides a pair of output signals for time-series analysis, using the harmonic filter (HF), while the bottom half gives a pair for power spectral analysis, after performing the power interpolation (PI).

The remainder of the power is left with the revised harmonic estimate $\tilde{V}(k)$, so we have:

$$\tilde{V}(k) = \begin{cases} \sqrt{1-\lambda(k)^2}\,\hat{V}(k) & \text{for } k \in B, \\ \hat{V}(k) & \text{otherwise;} \end{cases} \tag{5.14}$$

$$\tilde{U}(k) = \begin{cases} \hat{U}(k) + \lambda(k)\hat{V}(k) & \text{for } k \in B, \\ \hat{U}(k) & \text{otherwise.} \end{cases} \tag{5.15}$$

Hence, by using the original phase information for both components, $\arg(S_w(k))$, we can reconstruct the power-based time series $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$ in a way that is consistent from frame to frame. These signals retain the detail of the original time series, while avoiding misleading artefacts in the power spectrum in the form of troughs at the harmonics. Thus, the algorithm generates four complex spectra, $\hat{V}(k)$, $\hat{U}(k)$, $\tilde{V}(k)$ and $\tilde{U}(k)$, from a single input, as summarised in Table 5.1. After inverse-transforming and windowing, these are output as four time-series signals: $\hat{v}_w(n)$, $\hat{u}_w(n)$, $\tilde{v}_w(n)$ and $\tilde{u}_w(n)$, respectively. Each of these can be combined with the outputs from previous frames by sequential overlapping and adding to reconstruct two pairs of complete signals in alignment with the original signal $s(n)$: the harmonic and anharmonic signal estimates $\hat{v}(n)$ and $\hat{u}(n)$, and the harmonic and anharmonic power estimates $\tilde{v}(n)$ and $\tilde{u}(n)$.

		Voiced	Unvoiced
Signal	$k \in B$	$\hat{V}(k) = 2S_w(k)$	$\hat{U}(k) = S(k) - 2S_w(k)$
estimate	$k \notin B$	$\hat{V}(k) = 0$	$\hat{U}(k) = S(k)$
Power	$k \in B$	$\tilde{V}(k) = 2\sqrt{1-\lambda(k)^2}\, S_w(k)$	$\tilde{U}(k) = S(k) + 2\,(\lambda(k) - 1)\, S_w(k)$
estimate	$k \notin B$	$\tilde{V}(k) = 0$	$\tilde{U}(k) = S(k)$

Table 5.1: Spectral estimates for signal and power quantities for harmonic ($k \in B$) and anharmonic ($k \notin B$) frequency bins.

5.3.5 Note on robustness

The robustness improvement from using a Hann window, compared to a rectangular window, can be described by the sensitivity of cross-term bias errors between harmonics to deviations from perfect periodicity. These errors are reduced by a factor of 15 by the Hann window (i.e., 24 dB) at the adjacent harmonic, four bins away, as shown in Figure 5.4. Also, the half-power bandwidth of the main peak is increased from 0.44 bins to 0.72 bins at each harmonic, an increase of 60 %, which is related to the consequent increase in estimation variance. Therefore, despite being based on a maximum likelihood approach for estimating harmonically-related sinusoids, some of the idealised performance has been compromised to make the process more suitable for time-varying signals and much more robust.

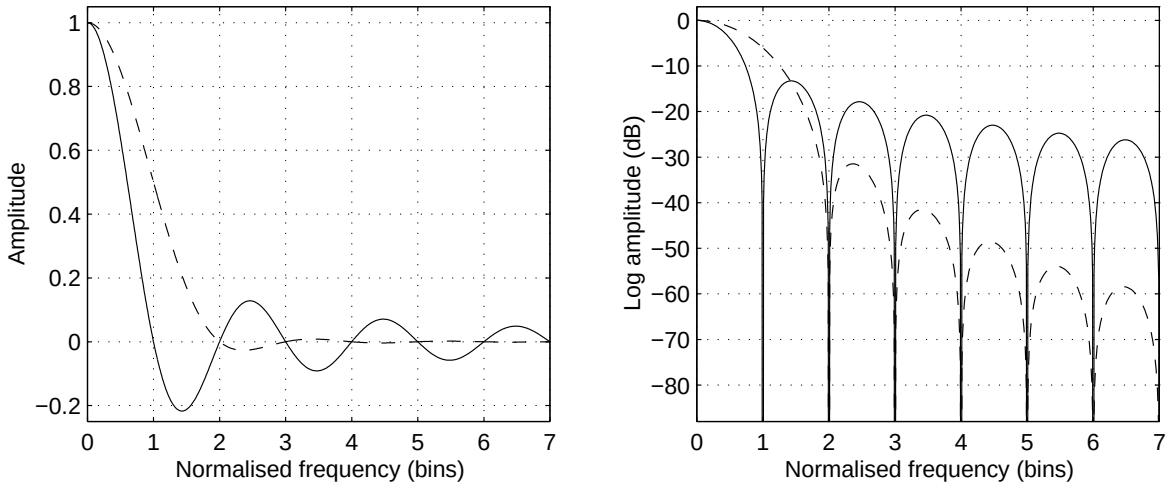


Figure 5.4: Smearing effect of the rectangular (solid) and Hann (dashed) windows on the spectral envelope, plotted on linear (left) and log (right) amplitude scales.

5.4 Selected methods

All the methods that were chosen for the pilot study assume that the speech signal $s(n)$ is the sum of a voiced component $v(n)$ and an unvoiced component $u(n)$, thus: $s = v + u$. They

attempt to separate the signal components by parametrically modelling the voicing to form an estimate $\hat{v}(n)$, which can be subtracted from $s(n)$ to yield an estimate of the unvoiced component $\hat{u}(n)$. In the rest of the section, we give a brief description of three alternatives to the PSHF. All four methods are then tested in the following section and their performances assessed.

5.4.1 Comb filter

Comb filtering is a time-domain technique that is well-suited to the enhancement of periodic signals. It can be envisaged as an average of equally-spaced points in the time domain synchronised to the underlying cyclical process or, in the FD, as a filter that has a spectrum of periodic spikes, rather like the teeth of a rake or a comb (hence the name). In speech, the spacing of the points is matched to the occurrence of glottal pulses. Hence the filter's frequency response is designed to coincide with the pitch harmonics.

By averaging periodically, the comb filter reduces the undesired contribution of the aperiodic part, which has zero mean and therefore tends to zero as the number of points M in the average tends to infinity:

$$\hat{v}(n) = \frac{1}{M} \sum_{m=0}^{M-1} s\left(n + m\,T_0\right) \ . \tag{5.16}$$

Thus, for synthetic signals, we can use a priori knowledge of the pitch period T_0 to yield an ideal averaging function by specifying the spacing of the teeth of the comb, but for real speech, either a pitch estimation process or an independent measurement is needed to provide T_0 values. However, errors in the timing of pitch pulses produce disproportionately large errors in the output, especially at the higher harmonics where phase differences from a time offset are magnified.

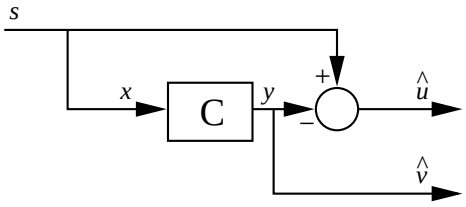


Figure 5.5: Comb filter (C) architecture, showing how the input speech s is decomposed into periodic and aperiodic signals, $\hat{v}$ and $\hat{u}$ respectively.

The method, depicted in Figure 5.5, is equivalent to segmenting the signal into pitch periods, aligning them to form an ensemble, and averaging across the ensemble, to produce a single segment, each point of which is the average of M points ($M = 100$, in the simulation). The single segment is then replicated, resulting in a periodic waveform of the same length as the

input signal. The estimate of the unvoiced component is calculated as the difference between this periodic estimate of the voiced component and the original input signal.

Adaptive comb filter

The adaptive comb filter differs from the standard comb filter in two important respects: (i) the spacing of points can be adjusted for variation in the pitch, and (ii) the points can be weighted. Where the pitch period, the time between two successive glottal pulses, is above average, the points can be placed further apart, and conversely. This leads to the problem of how to align periods of differing lengths, which has been resolved in the past either by truncation or zero-padding, (Shields 1970; Frazier et al. 1976; Lim et al. 1978). There are other possibilities with better properties that have not been explored, including the extrapolation of short periods with a parametric model of the waveform for that period.

In an M-point average, an even weighting is given to each point, which is equivalent to putting a boxcar window over the pitch periods. Better time resolution can be achieved by reducing M (at the expense of filter performance), but shaping the weights can improve adaptation, avoid artefacts from sidelobes, and help by emphasising the most relevant information. An example of a 5-point Hamming weighting is shown in Figure 5.6.

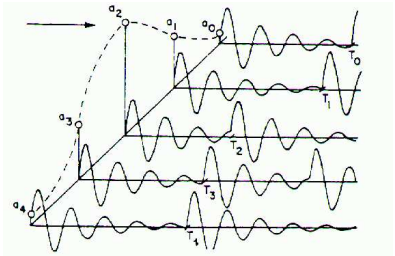


Figure 5.6: From Frazier et al. (1976), describing the operation of an adaptive comb filter.

5.4.2 Wiener filter

The Wiener filter was selected for preliminary testing because it is optimal for a broader class of signals, which are not completely independent. The typical Wiener filter architecture compares the input signal $s(n)$ with a filtered version of the reference signal $y(n)$. The output of the filter,

$$\hat{v} = W * y\,, \tag{5.17}$$

is the least-squares estimate of the input signal, which is formed from the inverse of the auto-correlation matrix R_{yy} and the cross-correlation vector p_{ys}:

$$W = R_{yy}^{-1} p_{ys}\,. \tag{5.18}$$

In our case, the reference signal was a time-shifted copy of the input signal: $y(n) = s\,(n + T_0)$, where the delay was chosen to match the pitch period T_0, as shown in Figure 5.7. The correlations were calculated using the entire record (1 s), which was wrapped to remove any end effects. An additional filter parameter, the filter order, was chosen to equal the number of samples of the delay, since it provided the maximum number of degrees of freedom allowed.

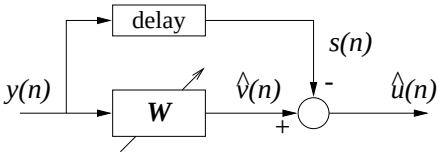


Figure 5.7: Wiener filter (W) architecture, which uses adaptive estimates of the auto-correlation and cross-correlation to predict the deterministic signal $\hat{v}$, and the residual $\hat{u}$ from s.

5.4.3 Thresholded wavelet filter

In the wavelet filter, the voiced component is modelled by a number of wavelets that is limited by a thresholding operation. The discrete wavelet transform of the incoming signal $s(n)$ is computed, and a threshold is applied to its coefficients, as illustrated in Figure 5.8. Those falling

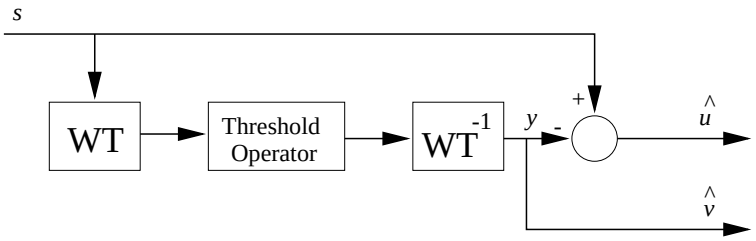


Figure 5.8: Wavelet filter architecture, containing the wavelet transform (WT), the thresholding operation and the inverse transform (WT^{-1}).

below the threshold are shrunk toward zero, then the inverse wavelet transform is computed to yield the estimated voiced signal, $\hat{v}(n)$. The non-linear operation of thresholding is designed to remove noise from a signal since the desired signal (the harmonic part, in this case) has its energy concentrated into a few wavelet coefficients, while the noise is more distributed across the wavelet space (Donoho 1993). The Daubechie 16 wavelet was chosen since it appeared to give the best results, according to visual assessment, from the choice of wavelet types offered by the software package, `xwpl-1.3`, obtained from Yale University (Majid 1997). No attempt was made to optimise the level of the threshold.

5.4.4 Discussion

All these methods are forms of averages, and they are all ways of sampling time-frequency space. They are all optimal in some sense, but each has its own short-comings. So, it is advisable to try to match the choice of method to the features for which one is interested. The comb filter performs an obvious arithmetic average, resulting in a biased power spectrum for the higher harmonics, where the true periodic component is below the level of the noise. Any errors in pulse timing create problems for the decomposition and can easily cause spurious outputs that have a metallic sound to them, because of the greater distortion of high-frequency components. Asynchronous methods give performance that depends on the number of periods in the window. Provided f_0 is unbiased, a reasonable separation can be achieved, depending also on how the bins are chosen, e.g., the cepstral method allocates half the bins to the periodic component (de Krom 1993; Gabelman et al. 1998; Yegnanarayana et al. 1998). However, asynchronous methods are less accurate than the PSHF, although computationally more efficient. The Wiener filter and other kinds of least-squares approaches effectively use a kind of threshold, depending on the number of points in the inverse filter. One drawback is that sometimes there is little separation of R_{ss} and R_{yy} in time (or frequency), which creates problems for the matrix inversion because of the low rank. Using a single delay to predict the deterministic component produces a sort of two-point comb filter, albeit a least-squares one. For the wavelet filter, the thresholding operation assumes that the deterministic structure is captured by a few wavelet coefficients, while the noise is distributed over the wavelet time-frequency space, whose coefficients are therefore lower in amplitude. The validity of these assumptions is strongly dependent on the HNR, the form of the voiced and unvoiced signals, and the wavelet base functions themselves.

The PSHF averages too. It provides both time-domain and frequency domain outputs effectively, thus giving scope to remove any bias. The concentration of energy is more efficient than asynchronous methods, which yields performance benefits. However, the computational complexity of the current algorithm makes it better suited to analysis than to coding applications. There are at least two possible ways of extending the technique: (i) using an assumption about higher order statistics (not generally well-defined for speech signals), (ii) using probabilistic models of speech components to improve accuracy of maximum likelihood estimates.

5.5 Comparative study

The purpose of the signal decomposition simulations was to compare the performance of the selected techniques against that of the PSHF. The approach has been to synthesise a signal from a combination of periodic glottal pulses and white noise, to apply various filtering techniques

attempting to reconstruct both the 'voiced' (harmonic) and 'unvoiced' (anharmonic) parts, and then to evaluate their performance in an objective way.

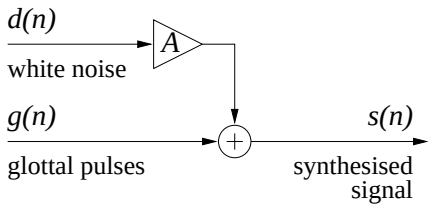


Figure 5.9: Schematic of basic signal synthesis model using glottal pulses, $g(n)$, and white noise, $d(n)$, which is amplified by the gain factor A.

5.5.1 Basic model

Using a simply generated waveform, containing idealised glottal pulses, $g(n)$, and Gaussian white noise, $d(n)$, various filtering techniques were tried in order to recreate the harmonic and anharmonic components, $v(n) = g(n)$ and $u(n) = A\,d(n)$ respectively (see Figure 5.9). The glottal pulse were constructed from a cubic function, as described by Klatt (1987), using the following parameters: the flow offset during the 'closed' phase ($20\,\mathrm{cm}^3/\mathrm{s}$), peak flow ($500\,\mathrm{cm}^3/\mathrm{s}$), fundamental period ($10\,\mathrm{ms}$, $f_0 = 100\,\mathrm{Hz}$), pulse duration ($5\,\mathrm{ms}$, $\mathrm{OQ} = 0.5$). A hundred pitch periods ($1\,\mathrm{s}$) of signal were generated at a sample rate of $f_s = 8.2\,\mathrm{kHz}$. The noise source was produced by a pseudo-random number generator with a normal distribution of unit standard deviation and zero mean. The gain, $A = 50$, was chosen arbitrarily and resulted in an initial HNR of $13.4\,\mathrm{dB}$. For each component and their sum, Figure 5.10 shows (left) one pitch period of time series, and (right) the power spectra.

5.5.2 Performance calculation

From decomposition of the speech $s(n)$, we want a harmonic signal $\hat{v}(n)$ that represents the best estimate of the voiced component, defined as having the minimum mean-squared error between the actual voiced component time series $v(n)$ and the estimate $\hat{v}(n)$. Similarly, we want the anharmonic estimate $\hat{u}(n)$ to be as near to the additive noise $u(n)$ as possible. Since $\hat{v} + \hat{u} = s$, the error, defined as $e(n) = \hat{v} - v = -(\hat{u} - u)$, is equal and opposite in the harmonic and anharmonic components.

The performance of the PSHF was assessed by considering the change in signal-to-error ratio (SER) for each component. The jitter and shimmer perturbations of the pulse train were considered intrinsic to the synthetic voicing signal, whereas the additive noise was treated as the product of another source, representing the unvoiced component. Therefore, for the harmonic component, the additive noise was the initial 'error' on the voiced 'signal' component. Conversely, for the anharmonic part, the voiced component was taken to be the 'error'

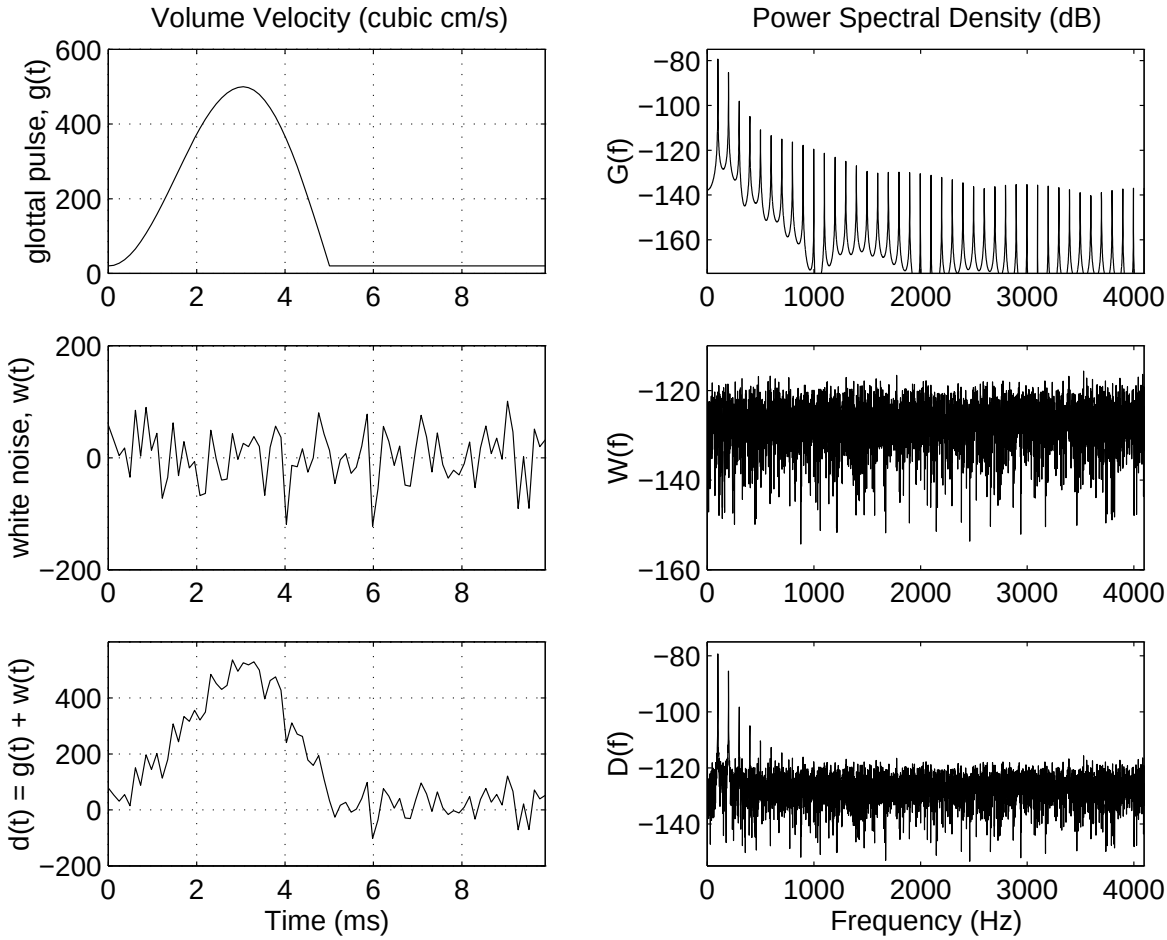


Figure 5.10: Time series (left) and power spectra (right) of periodic and aperiodic components, and their sum: (top) one glottal pulse, $v(n) = g(n)$, (middle) white noise, $u(n) = A\,d(n)$, and (bottom) the combined signal, $s(n) = v(n) + u(n)$. The power spectra, $V(k)$, $U(k)$ and $S(k)$, were calculated from the entire sample (1 s).

of the unvoiced 'signal' initially. Hence, the harmonic performance η_v and the anharmonic performance η_u (expressed in decibels) are:

$$\eta_v = 10\log_{10}\left(\frac{\langle v^2\rangle/\langle e^2\rangle}{\langle v^2\rangle/\langle u^2\rangle}\right) = 10\log_{10}\left(\frac{\langle u^2\rangle}{\langle e^2\rangle}\right), \text{ and} \tag{5.19}$$

$$\eta_u = 10\log_{10}\left(\frac{\langle v^2\rangle}{\langle e^2\rangle}\right). \tag{5.20}$$

Although these two expressions are clearly related by the HNR σ_N (i.e., $\eta_u = \sigma_N + \eta_v$), it is useful to describe the performance of both components separately.

Basing our performance on the MSE may give results for time-varying signals that do not necessarily correspond to speech intelligibility (Lim et al. 1978). Nevertheless, it is highly desirable to have such a common currency as an objective measure of the performance, providing a solid scientific basis for evaluation of the various techniques. It follows that evaluating the change in SER for the periodic and aperiodic estimates from the synthetic speech constitutes a more rigorous performance metric for reconstructing signals than a comparison of prescribed HNR (before synthesis) versus measured HNR (after decomposition). So, although we include some HNR measurements to aid comparison with other algorithms, we generally use the SER to describe the performance of the PSHF.

Alternative metrics

Although we require performance measures for assessing the filters with synthetic signals, these measures cannot generally be applied to real speech. Therefore, some alternative measures that can be applied to real speech would be useful. There are also other factors that may be important for particular applications of the decomposition filters, such as speed in real-time speech coding.

Among alternative metrics there is the mean error magnitude ('city-block' distance), which can be used to minimise the error signal, rather than its power. The Chebyshev metric (Deller et al. 1993), which uses the maximum error, tends to reduce the number of outliers, which could help to highlight impulsive elements of the signals. The mean cepstral error (from comparison of the log-spectra), by giving equal significance to the zeros as to the poles of the transfer function, is likely to be beneficial for studying non-glottal sources.

There are also ways of trying to relate the signals to their perceptual properties, using A-weighted spectra, mel-spaced frequencies, and critical bands (e.g., in terms of Barks). The ultimate method of assessment however is through perceptual testing, but these kinds of tests are not trivial to conduct and require a significant number of participants to make the assessment statistically valid.

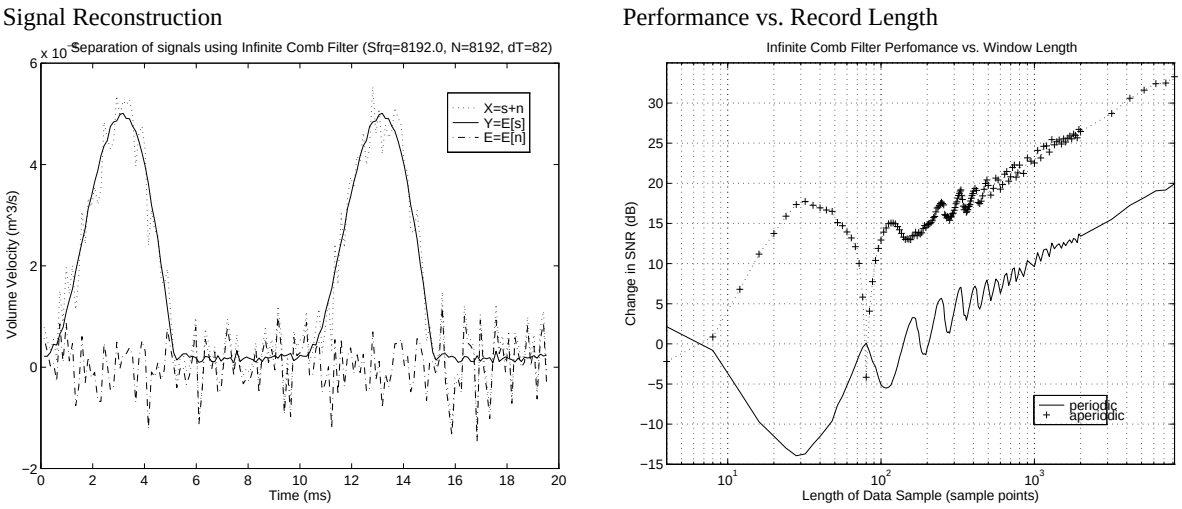


Figure 5.11: Left: signal reconstruction applying the infinite comb filter to the full record (1 s), showing the input signal, s (dotted), the harmonic estimate, $\hat{v}$ (solid), and the anharmonic estimate, $\hat{u}$ (dash-dot). Right: variation of filter performance against the length of record, harmonic (solid) and anharmonic (crosses).

5.5.3 Comb filter

The comb filter increased the SER of the harmonic component from 13.4 dB to 33.3 dB (i.e., $\eta_v = 19.9$ dB), and the anharmonic component from -13.4 dB to 19.9 dB ($\eta_u = 33.3$ dB). Figure 5.11 (left) shows the reconstruction of the harmonic and anharmonic signals from the mixed input signal, and (right) a plot of the variation of filter performance against the length of record. The filter started to obtain positive performance on the periodic component once there were almost two complete periods (164 samples) in the record over which to average. The components' performance results for shorter record lengths are somewhat misleading and actually in antiphase to each other. However, as the record length was increased, peaks in the performance on both components was observed at integer numbers of the period: 246, 328, 410 samples, etc. The overall trend increased at 10 dB/decade.

5.5.4 Wiener filter

The simple time-delay architecture of the Wiener filter may give good results, but its effect on the anharmonic spectrum is not well understood at this stage, and so interpretation of the results must proceed with caution. This filter increased the SER of the harmonic component from 13.4 dB to 22.1 dB ($\eta_v = 8.8$ dB), and the anharmonic component from -13.4 dB to 9.3 dB ($\eta_u = 22.7$ dB). The reconstruction with zero initialisation (Figure 5.12, left) shows a transition stage during the first pitch period (0–10 ms) as the internal states of the filter (initially at zero) are adjusted. This effect can be avoided by allowing the filter to consider the signals to be

periodic over the sample length, and wrapping them (right). Note how the latter part (> 10 ms) of the plotted response remains unaffected.

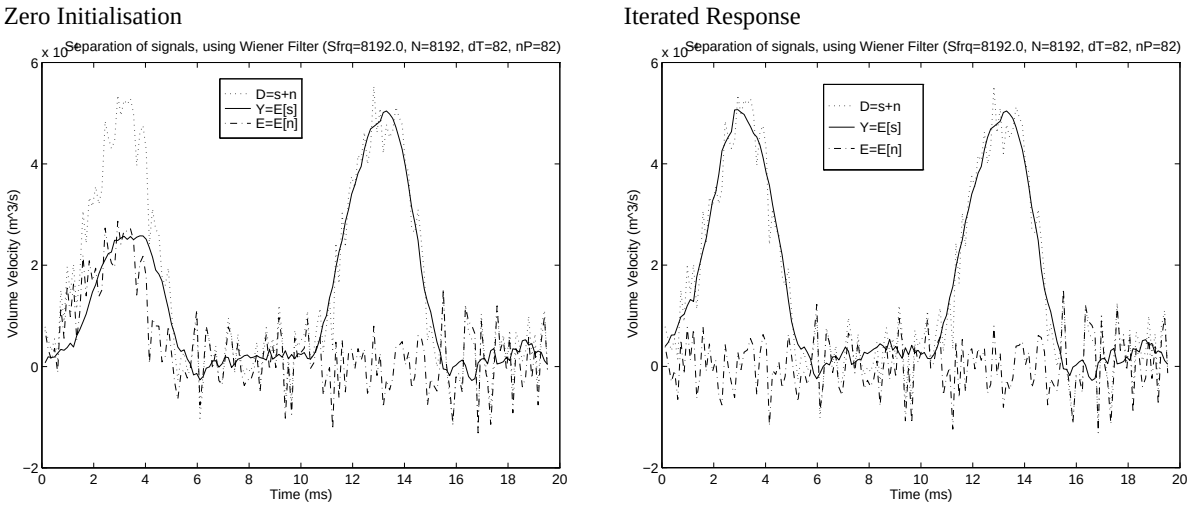


Figure 5.12: Signal reconstruction using Wiener filter, with zero initialisation (left) and iterative (repeated-sample) initialisation (right), showing the input signal, s (dotted), the harmonic estimate, $\hat{v}$ (solid), and the anharmonic estimate, $\hat{u}$ (dash-dot).

The poor performance of the Wiener filter at discontinuities can be seen in the residue signal, which 'blips' periodically at each glottal closure (at 5 ms, 15 ms, etc.). The periodicity is reflected in the spectrum of the residue, which has energy at the third and higher harmonics of the fundamental frequency (the first two have been attenuated).

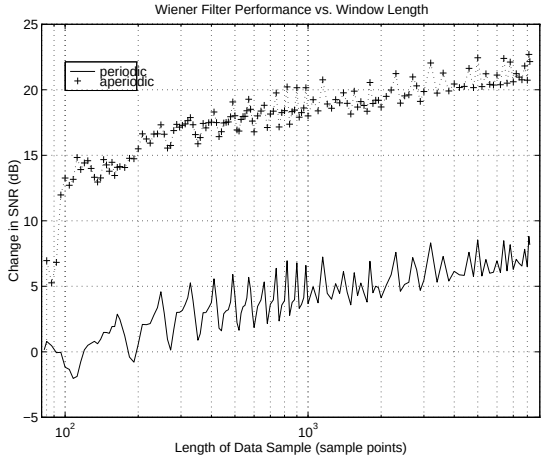


Figure 5.13: Performance of the Wiener filter against filter length, harmonic (solid) and anharmonic (crosses).

The variation of filter performance against the record length is plotted in Figure 5.13, which shows ripples, similar to those seen previously in Figure 5.11 (right), that are an effect of the scaling of the record length to an integer number of pitch periods. The asymptotic

performances rise at only 5 dB/decade for the Wiener filter.

5.5.5 Wavelet filter

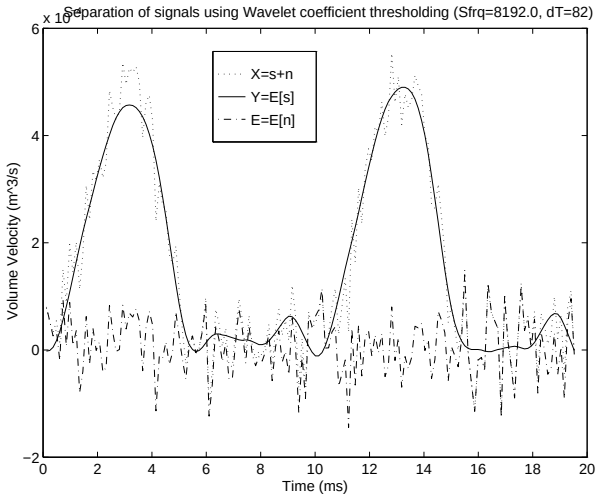


Figure 5.14: Signal reconstruction using thresholding of Daubechie (16) wavelet coefficients, showing the input signal, s (dotted), the estimate of harmonic component, $\hat{v}$ (solid), and the estimate of the anharmonic component, $\hat{u}$ (dash-dot).

The results of preliminary tests using the Daubechie 16 wavelet are shown in Figure 5.14. Note that the estimate of the harmonic component $\hat{v}$ is not itself periodic. The effective removal of some of the smaller, higher frequency terms has created a smoother waveform, which has difficulty coping with the discontinuities at the start (at 10 ms) and finish (at 5 ms) of the glottal pulse. It may be possible to improve the results by experimenting with alternative wavelet functions. Neither has the thresholding operation been optimised for this signal processing problem. Nevertheless, the wavelet filter increased the SER of the harmonic component from 13.4 dB to 19.3 dB ($\eta_v = 5.9$ dB), and the anharmonic component from -13.4 dB to 6.2 dB ($\eta_u = 19.6$ dB).

5.5.6 Pitch-scaled harmonic filter

The PSHF increased the SER of the harmonic component from 13.4 dB to 17.6 dB ($\eta_v = 4.2$ dB), and the anharmonic component from -13.4 dB to -11.2 dB ($\eta_u = 2.2$ dB). As can be seen in Figure 5.15, the PSHF operates on a short windowed section of the signal that is four pitch periods long, unlike the other selected techniques. The outputs of successive frames are typically accumulated to process long continuous sections.

The example illustrated here shows corresponding windowed outputs: the harmonic estimate with a reduced noise level, and a noisy anharmonic estimate without any major excursions from the origin. Notice how the PSHF seeks to assign the very low-frequency part of the input

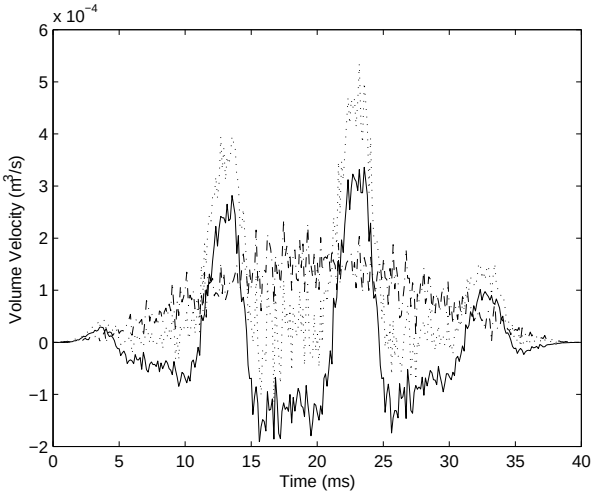


Figure 5.15: Partial signal reconstruction using PSHF (over four pitch-periods, 40 ms), showing the input signal, s (dotted), the harmonic estimate, $\hat{v}$ (solid), and the anharmonic estimate, $\hat{u}$ (dash-dot).

signal ($< f_0/2$) to the anharmonic component. Because of the window, this is manifested as an arching noise signal, while the harmonic component oscillates evenly about the origin. This effect produces abnormally low performance results since a d.c. component is not usually present in sound recordings. However, by compensating for the zero offset we obtain $\eta_v = 5.3\,\mathrm{dB}$ and $\eta_u = 18.4\,\mathrm{dB}$ for the harmonic and anharmonic performance, respectively.

5.5.7 Pilot summary

Table 5.2 summarises the performance results of the selected filters, which indicate that the infinite comb filter has been the most successful at separating the two sources. Using the full length of the 1 s record to estimate the periodic part of the signal clearly yields better results than using just a short section of the signal, here 40 ms long. This is evident for both methods that offer comparison, much as would be expected for any sort of averaging process. The extraction of the deterministic part by the wavelet method gave much poorer results on the full record, relative to the other techniques, and might be expected to give correspondingly low performance on the partial record. Since the PSHF only uses a much shorter section of the input signal to make its estimates, its performance falls short of that achieved by the other techniques on the full record. When tested on frames of equivalent duration, however, the values are comparable.

Although the PSHF is not the best-performing technique in this highly idealised simulation, coming second to the infinite comb filter, its performance in the partial record test is similar to that of the comb and Wiener filter methods, and to the full-record result for the wavelet method. This is encouraging because the PSHF has been designed for processing realistic speech-like

Technique	Change in SER (dB)	
	Full Record (1 s)	Partial Record (40 ms)
	η_v, η_u	η_v, η_u
comb	19.9, 33.3	7.0, 19.0
Wiener	8.8, 22.7	5.3, 17.9
wavelet	5.9, 19.6	—, —
PSHF	—, —	5.3, 18.4

Table 5.2: Performance of filters, η_v and η_u, on synthesised signal with no jitter, no shimmer and initial HNR = 13.4 dB.

signals that vary in many ways over time, rather than for a perfectly periodic infinite signal in noise. In particular, we would expect the comb filter to suffer a severe performance degradation in the presence of even quite mild jitter and shimmer, and hence to reverse the pecking order.

Apart from wanting to pursue the PSHF for the sake of evaluating this novel technique, it has the advantage of not requiring precise pitch epochs to be determined (in contrast to the comb filter). For the study of voiced fricatives and other examples of weak voicing, such as breathy vowels, where these epochs are less well defined and identified less reliably, removing this requirement is especially important for achieving a faithful decomposition. Thus, despite the moderate performance of the PSHF, we are reassured that it shows promise for equitable results on application to real speech. A better trial of the techniques would therefore involve a changing quasi-periodic source instead of the stationary example used here. In the following chapter, we show the result of decomposing real speech with the PSHF, which has the most readily interpretable output, in terms of the speech spectra, but first its performance is assessed under less ideal conditions, to give an indication of its actual effectiveness in practice.

5.6 Validation using synthetic speech

Speech-like signals were synthesised to test the PSHF algorithm for any processing artefacts, and to evaluate its performance. The signals were generated using a TD method to avoid the possibility of spurious results caused by the interaction of two FD methods, as might occur when synthesis frames are synchronised with analysis frames. The purpose of decomposing the signal is to produce accurate representations of the individual components. Ideally, we would like to reconstruct them perfectly, so that they could be analysed, alongside other parameters (e.g., mean flow rate, vocal effort, f_0, sex of speaker, etc.), without interference from each other. Therefore, the decomposed synthetic signals were evaluated for the filters' ability to reduce the interference. Real speech signals suffer from fluctuations in amplitude and frequency of the

glottal pulses (i.e., jitter and shimmer), and so the proposed filter technique was also assessed under such conditions.

5.6.1 Signal generation

The PSHF was tested with synthetic speech-like signals and the accuracy of its decomposition evaluated. The signals $s(n)$ were generated in the TD (avoiding any potential artefacts from later FD filtering) by convolving excitation signals $c(n)$ with an appropriate filter $q(n)$:

$$s(n) = c(n) * q(n)\,. \tag{5.21}$$

Each excitation signal $c(n)$ was the sum of a GWN signal $d(n)$ and a pulse train $g(n)$ with sample values $\{\ldots, 0, 1, 0, 0, \ldots\}$:

$$c(n) = g(n) + d(n)\,. \tag{5.22}$$

In the first series of tests, the pulse train was periodic. In some cases, the amplitude of the noise $d(n)$ was modulated in time with the pulses. Thus, the noise was combined in three ways: (i) with constant-variance; (ii) with the amplitude of the noise modulated by a sinusoid at the fundamental frequency f_0, in anti-phase with the glottal excitation,

$$u = A\,(d * q)\,\sqrt{\frac{2}{3}}\left[1 + \cos\left(\frac{2\pi f_0 n}{f_s} + \beta\right)\right]\,, \tag{5.23}$$

where $\beta = \pi$; (iii) modulated by a rectangular wave to give a 60 % burst duration with respect to the pitch period. The factors of $\sqrt{2/3}$ and $\sqrt{1/0.6}$ compensated for the effects of the modulation on mean signal power in (ii) and (iii), respectively. In all cases, the gain of the noise signal A was adjusted relative to that of the pulse train to give HNRs at one of six specified levels: $\sigma_N \in \{\infty, 20, 10, 5, 0, -5\}$ dB.

A set of linear predictive coding coefficients (LPC, 50-pole autocorrelation) was computed for a male [ɑ], using a section from the middle of the first vowel in a recorded nonsense word (see Section 6.2.1 for details). Each excitation signal, $c(n)$, was passed through the corresponding LPC synthesis filter, $q(n)$, at sampling rate of 48 kHz.

5.6.2 Results

First, the cost function $J(N, p)$ was used by the pitch tracker ($H = 8$ harmonics) to optimise the window length $N(p)$ for each synthetic signal. Using the specified average fundamental frequency $\bar{f}_0$ to give an initial estimate $N_{\text{init}} = 4/\bar{f}_0$, the local minimum in J was found at a series of points p throughout each test signal. For high HNRs, the estimated period was identical to the true T_0 but, as the noise level was increased, the deviation of the estimates also increased. These values were given as the pitch input to the PSHF, which then decomposed the signals into harmonic and anharmonic components, $\hat{v}$ and $\hat{u}$ respectively, the estimates of the voiced and

unvoiced parts. Each signal was processed in the usual way: incrementing the analysis frame, decomposing and accumulating the outputs. For this study, we were deliberately conservative, centring frames on every sample point (offset $\alpha = 1$), which was computationally expensive.

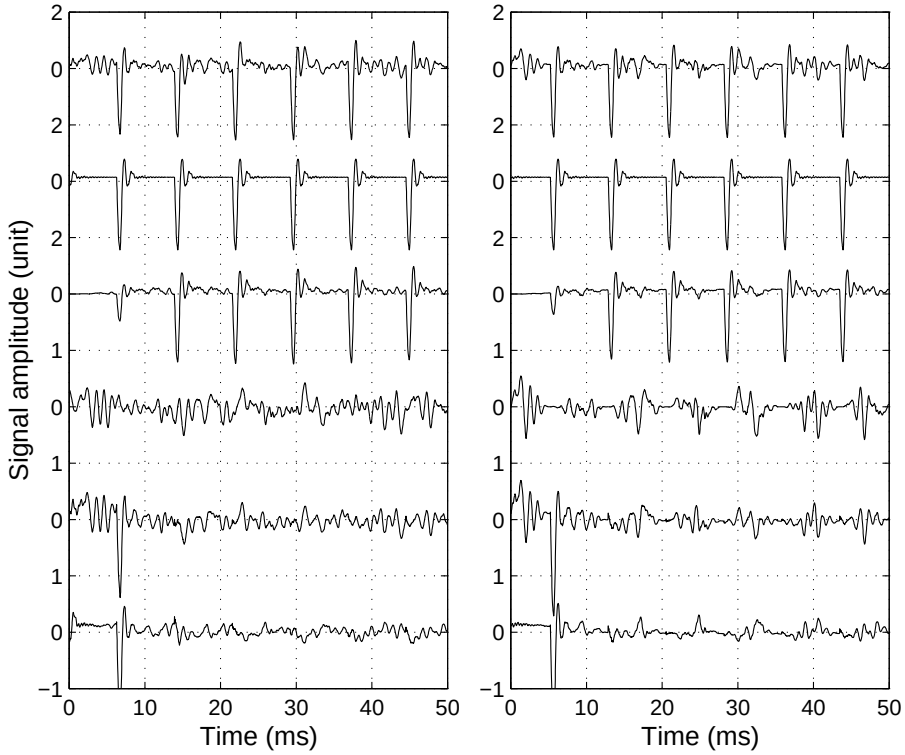


Figure 5.16: Time series of the synthetic signal $s(n)$ with its constituent harmonic and anharmonic parts $v(n)$ and $u(n)$, the PSHF signal estimates $\hat{v}(n)$ and $\hat{u}(n)$, and the error $e(n)$, at HNR = 10 dB for (left) constant-variance noise, and (right) modulated noise with $\beta = \pi$. They are arranged, from top to bottom, thus: s, v, $\hat{v}$, u, $\hat{u}$ and e (anharmonic and error signals are double amplitude scale).

Although there are transient errors for the first two pitch periods, as the tail of the first window ramps up towards its centre, the decomposed components shown in Figure 5.16 soon approach the true components. Looking at the time series more closely, it is apparent that the modulation of the noise envelope is retained. Indeed, the error signal also exhibits some modulation, suggesting that the error is proportionally related to the noise, for a given mean HNR. The amplitude of the envelope of $\hat{u}$ is slightly reduced with respect to the input component u, but *its phase remains unaltered*. This finding, which is crucial to the results presented in Chapter 7, will be further justified Section 7.2.2. These simulations, therefore, support the assertion that any modulation exhibited by the anharmonic component is not a processing artefact, but a property of the source component from which it is derived.

Figure 5.17 illustrates the analyses of two examples: (a) with constant noise and (b) with modulated noise. In each of the figures, the top curve is the synthesised signal $s(n)$ that was

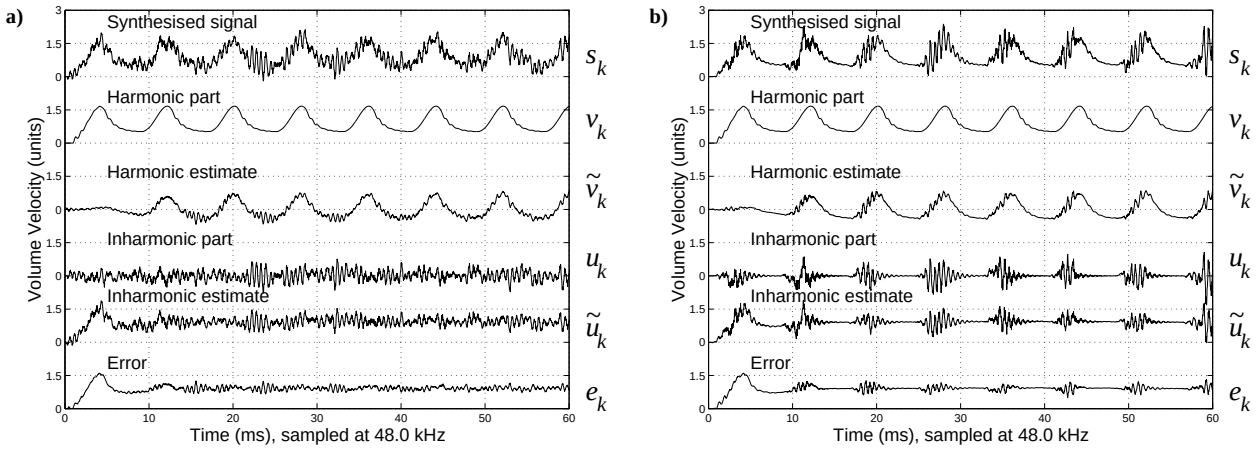


Figure 5.17: Time series of synthetic signals with their constituent harmonic and noise parts, and the respective PSHF estimates, with (a) constant noise and (b) modulated noise: (from top) s, v, $\hat{v}$, u, $\hat{u}$, and the error e.

fed into the PSHF. The second and fourth curves are the harmonic part $v(n)$ and anharmonic part $u(n)$ respectively, that generated it: $s = v + u$. The third and fifth curves are the corresponding estimates, $\hat{v}(n)$ and $\hat{u}(n)$ respectively, given by the filter. The bottom curve is the error $e(n)$, which is equal and opposite for the two components: $e = \hat{v} - v = -(\hat{u} - u)$.

Note how the envelope of the noise signals is preserved. In Figure 5.17a, the anharmonic estimate $\hat{u}$ delivered by the PSHF is a constant amplitude noise signal, just like the input u; meanwhile, in Figure 5.17b the anharmonic estimate is modulated mimicking the modulation of the filtered noise input. In the latter example, Fig. 5.17b, the PSHF improved the signal-to-error ratio of the anharmonic part by 11.6 dB (from -5.9 dB to 5.7 dB).[2] In other words, in the synthesised signal, the modulated noise part was only of about half the amplitude of the periodic part, but in the extracted anharmonic estimate, the reconstructed noise signal was about twice the residual error.

5.6.3 Evaluation

The performance was evaluated in terms of the change in SER for each of the test signals. Table 5.3 lists the harmonic and anharmonic performances, η_v and η_u, over the range of specified noise conditions. Except for the anharmonic performance at the -5 dB condition, all the performance values are positive, which implies that the quality of the separated component

[2]Normal microphone recordings have a high-pass response that admits only an a.c. signal. To include flow-induced noise between the microphone's roll-on frequency and the fundamental frequency f_0, the PSHF has been designed to assign the bins below the first harmonic ($< f_0$) to the anharmonic component. Therefore, in this case, where the harmonic part $v(n)$ contained an offset, the mean was subtracted before calculation of the error. The transients at the beginning and the end of the filter outputs were also excluded from the calculation.

	Initial harmonics-to-noise ratio (dB)					
Noise type	∞	20	10	5	0	-5
constant	–, 72.6	5.3, 25.2	5.2, 15.1	5.2, 10.1	5.1, 4.9	5.1, -0.0
modulated	–, 72.6	5.6, 25.4	5.5, 15.4	5.4, 10.2	5.2, 5.0	4.2, -1.0

Table 5.3: PSHF performance versus HNR for synthetic signals with constant and modulated noise ($\beta = \pi$); results are η_u, η_v in dB.

is better than the input signal, i.e., the remaining errors are always smaller than the original corruption from the interfering source, for non-negative HNRs. The anharmonic performance is a strong function of HNR, and is approximately 5 dB greater than the initial HNR, so that any residual errors in the extracted unvoiced estimate are about half as large as the true unvoiced component. Meanwhile, the harmonic component is cleaned up to a similar degree by the PSHF, which reduces the errors to about half of their original amplitude, on average. Note that the results of the constant-variance noise case and modulated noise case ($\beta = 180°$) are almost identical, which implies that the performance is not significantly affected by the envelope of the noise. Tests at other phase settings produced similar results ± 0.2 dB. Overall, the results indicate the extent to which we can have confidence in the output signals that the PSHF produces.

Figure 5.18 shows the results for three periodic signals corrupted by various levels of either constant or modulated noise. The performance was positive in all but a few extreme cases, and was typically $\eta_v \approx 5$ dB for the harmonic component and $\eta_u \approx \sigma_N + 5$ dB for the anharmonic one. Thus, for a normal vowel with an HNR of 15 dB, the harmonic performance would be greater than 5 dB and the anharmonic performance approximately 20 dB (Awan and Frenkel 1994). For $\sigma_N \leq 0$ dB, the performance deteriorated and in some cases became negative; this deterioration was more pronounced for modulated noise. At infinite HNR ($\sigma_N = \infty$ dB), improvements in the anharmonic SER were 73, 54 and 50 dB respectively, for the three values of $\bar{f}_0$: 120, 130.8 and 200 Hz. Thus, pitch quantization and spectral smearing defined a performance limit by producing errors up to 1/300th of the original signal with no jitter, shimmer or noise disturbance present.

The results were almost identical for all $\bar{f}_0$ values, a characteristic of pitch scaling, except at low HNRs where pitch tracking errors produced spurious readings. Similarly, altering the envelope of the noise, although perhaps making the tracker more error-prone, did not significantly affect the quality of the decomposition. In our previous study (Jackson and Shadle 1998), signals with constant-amplitude noise and noise modulated by the glottal waveform were synthesised. Results of the decomposition showed that the respective constant and modulated envelopes of the reconstructed noise signals were retained, which suggests that any modulation

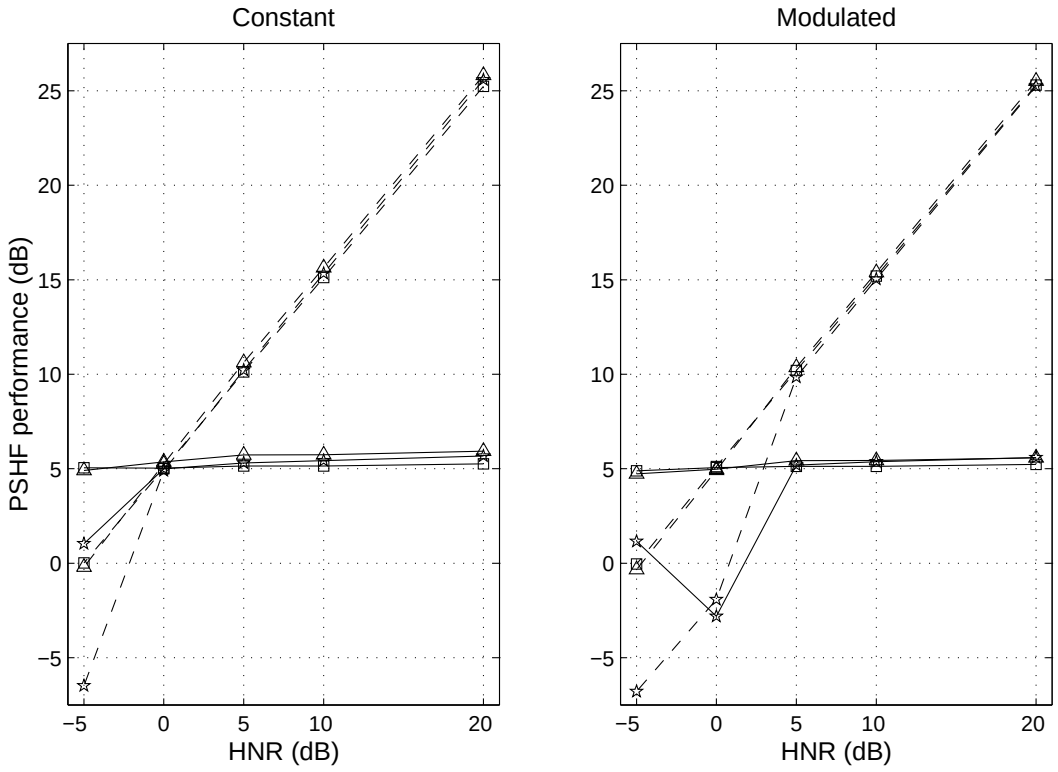


Figure 5.18: Anharmonic η_u (dashed) and harmonic η_v (solid) performance of the PSHF on synthetic speech signals versus HNR; with constant (left) and modulated (right) noise. Each graph shows results for three values of $\bar{f}_0$: 120 Hz (△), 130.8 Hz (star), and 200 Hz (box). No jitter or shimmer. See text for values at $\sigma_N = \infty$ dB.

observed in real speech is not a processing artefact.

Incidentally, repeating the process using the prescribed pitch values to determine $N(p)$ showed that using the noisy estimated values had little effect on the anharmonic performance, which was degraded by 0.4 dB in the worst case. The observed decline with increasing noise in the harmonic performance, though, was entirely due to the effect of noise on the estimated pitch, which would otherwise have kept η_v pinned at 5.3 dB and 5.6 dB for all constant and modulated noise tests, respectively.

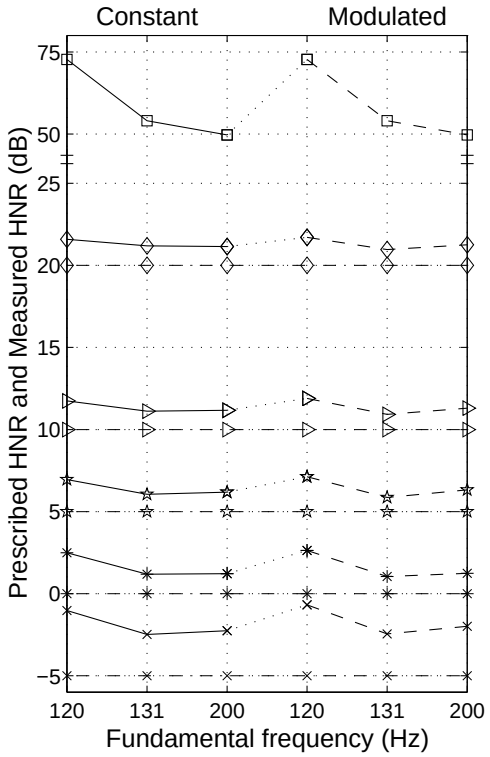


Figure 5.19: Measured HNR for constant (solid) and modulated (dashed) noise versus f_0, shown for the prescribed values (dash-dot, from bottom): −5 dB ($\times$), 0 dB ($*$), 5 dB (star), 10 dB ($\triangle$), 20 dB ($\Diamond$), ∞ dB (box, separate scale). No jitter or shimmer.

5.6.4 Measured HNR

Although not principally designed for such a purpose, the power-based outputs of the PSHF, $\tilde{v}$ and $\tilde{u}$, may be used as a measure of the total power of each component. Hence, by comparing $\langle \tilde{v}^2 \rangle$ with $\langle \tilde{u}^2 \rangle$, an estimate of the HNR may be formed, where $\langle\ \rangle$ denotes time averaging. The measured HNRs, calculated for the signals from Figure 5.18, are just above the true (prescribed) HNRs in all cases, except for $\sigma_N = \infty$ (the no-noise case), as shown in Figure 5.19. The measured HNRs varied little with $\bar{f}_0$, and the noise envelope (constant or modulated) had a negligible effect. The discrepancy between the measured and prescribed HNRs is largest for the cases with most tracking errors, i.e., at −5 dB, but otherwise it is c. 1–2 dB. Note that

the decomposition anomaly evident in Figure 5.18 ($\sigma_N = 0\,\text{dB}$ modulated, $\bar{f}_0 = 130.8\,\text{Hz}$) is not apparent in these results, because the measured HNR, which is the ratio of the component powers, is not based on the actual decomposed signals, but merely compares their mean-square values.

5.7 Effect of voicing perturbations

In the second experiment, forms of signal disturbance other than noise were introduced. Since the oscillating vocal folds often vary in timing and amplitude, these kinds of perturbation were added to the synthetic signals.

5.7.1 Signal generation

Test signals were made from voiced and unvoiced components:

$$\begin{aligned} s(n) &= v(n) + u(n) \\ &= [g(n) + d(n)] * q(n) \\ &= c(n) * q(n)\,, \end{aligned} \tag{5.24}$$

as before. Only constant-amplitude noise $d(n)$ was used, added at four levels with HNRs of ∞, 20, 10 or 5 dB. This time, however, the glottal pulse train $g(n)$ was modified. The pitch period (and hence f_0) and the amplitude of $g(n)$ were perturbed from their nominal values ($\bar{f}_0 = 130.8\,\text{Hz}$, $\bar{a} = 1$) by specified amounts of jitter (0, 0.25, 0.5, 1, 3 or 5 %) and shimmer (0, 0.5, 1.0 or 1.5 dB), respectively.[3] Normal values for jitter and shimmer during modal phonation are typically less than 0.7 % and 0.5 dB, respectively (Dworkin and Meleca 1997; less than 1 % and 0.25 dB according to Blomgren et al. 1998), although they can be as much as 3 % and 1 dB (Michaelis et al. 1995).

Jitter σ_T, specified as a percentage, was added to the synthetic signals by modifying the pitch-period epochs in the pulse train (Michaelis et al. 1995):

$$T_i = \frac{1}{\bar{f}_0}\left(1 + \frac{r_i\sqrt{\pi}}{2}\frac{\sigma_T}{100}\right), \tag{5.25}$$

where $\bar{f}_0$ is the nominal pitch frequency and r_i is a random variable with a Gaussian probability distribution of zero mean and unit standard deviation. The factor of $\sqrt{\pi}/2$ is needed to match

[3] These perturbations that we call jitter and shimmer do not necessarily represent realistic physical properties of f_0 variation, but are used to illustrate the effect of perturbations on the PSHF. The fine time resolution of the PSHF leaves it unaffected by low-frequency perturbations, such as vibrato, but the above test methodology provides quantitative and self-consistent results.

the standard deviation of T_i to the mean difference between two such variables, $|T_i - T_{i-1}|$. For shimmer, the pulse amplitude a_i was altered according to the expression:

$$a_i = \bar{a}\left(1 + \frac{r_i\sqrt{\pi}}{2}10^{0.05\,\sigma_A}\right), \tag{5.26}$$

where σ_A was the level of shimmer, specified in dB (Michaelis et al. 1995).

Using the set of linear predictive coding coefficients (LPC, 50-pole autocorrelation) computed for a male /ɑ/ to make $q(n)$, as before, each excitation signal, $c(n)$, was LPC filtered at the 48 kHz sampling rate. In evaluating the performance, the jitter and shimmer perturbations of the pulse train were considered intrinsic to the synthetic voicing signal, $v(n) = g(n) * q(n)$. Conversely, the additive noise was treated as the product of another source, corresponding to the unvoiced component, $u(n) = d(n) * q(n)$.

5.7.2 Results

The cost function $J(N, p)$ was used by the pitch tracker ($H = 8$ harmonics) to estimate $N(p)$ for each of the synthetic signals. Then the signals were decomposed by the PSHF algorithm into harmonic and anharmonic estimates, $\hat{v}$ and $\hat{u}$ respectively. Again for this experiment, we centred frames on every sample point (cautious but computationally-expensive offset $\alpha = 1$). The measured values of jitter and shimmer compared well with those specified.

Figure 5.20 illustrates the effects of jitter (left) and shimmer (right) on the PSHF performance, in combination with constant noise added at various levels. The trends are qualitatively similar for both perturbations. For example, when there is no noise, there is a notable performance degradation with the introduction of any jitter or shimmer. However, for the typical sort of variations observed in real speech (Michaelis et al. 1995; Blomgren et al. 1998), fluctuations in the pitch period (jitter) have a larger effect on performance than amplitude fluctuations (shimmer), which was also true over the range of values tested. Where there is already one disturbance, i.e., HNRs of 20, 10 or 5 dB, the introduction of a second one, either jitter or shimmer, is less marked. The performances are generally positive, except for η_v at the higher levels of jitter ($\sigma_T \geq 1.5\,\%$) and shimmer ($\sigma_A \geq 1.5\,\text{dB}$) with high HNR ($\sigma_N \geq 20\,\text{dB}$), for which the initial error was relatively small. Table 5.4 extends this principle to the combination of all three disturbances, whose worst element puts a bound on the performance. Indeed, the performance can even improve with additional perturbation, as occurred for jitter of 3 % when shimmer was added. For normal speech, the presence of all three disturbances degrades performance by 1 to 2 dB with respect to the noise-only case (i.e., Fig. 5.18).

In the presence of perturbation, the harmonic performance η_v generally improves as the noise level increases, although the quality of the final estimate is degraded in an absolute sense (i.e., the final error also increases). As before, the perturbations to the excitation signal, σ_T

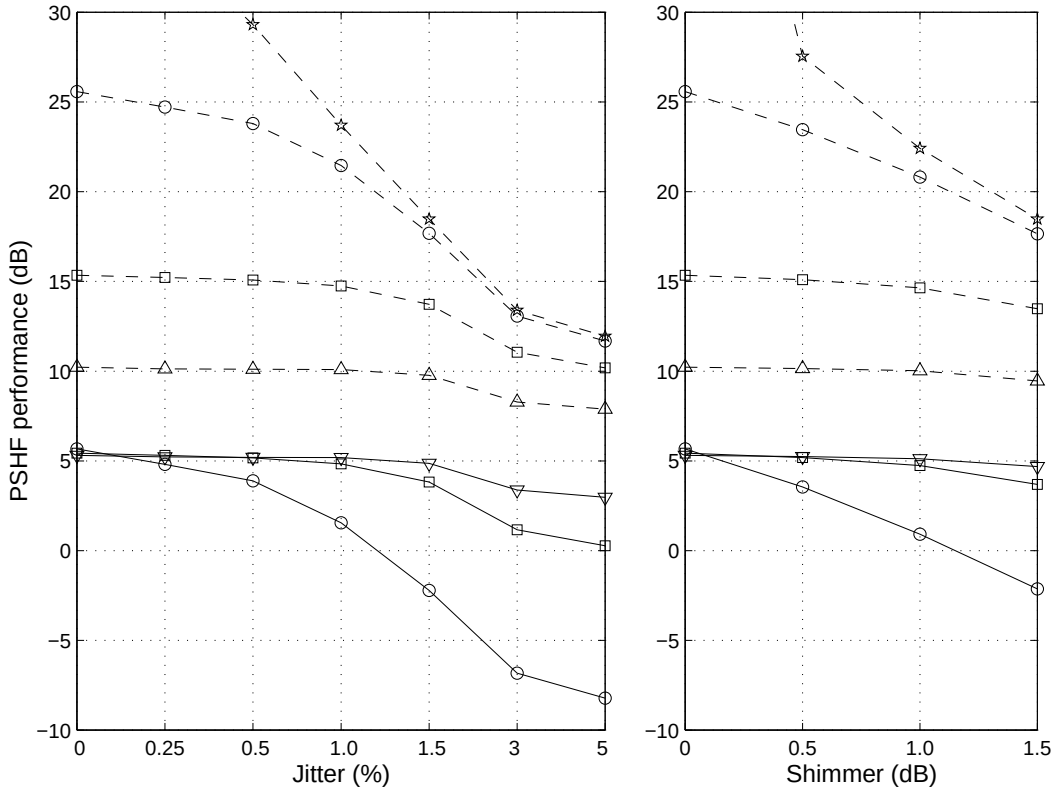


Figure 5.20: Anharmonic η_u (dashed) and harmonic η_v (solid) performance of the PSHF on synthetic speech signals, perturbed with either jitter (left) or shimmer (right). For both graphs, the HNRs are: ∞ dB (star), 20 dB (○), 10 dB (box), or 5 dB (△).

σ_T	σ_A	Initial harmonics-to-noise ratio, σ_N (dB)			
%	dB	∞	20	10	5
0	0	–, 54.0	5.6, 25.1	5.4, 14.9	5.3, 9.8
	1	–, 22.0	0.5, 19.9	4.7, 14.2	5.0, 9.6
0.5	0	–, 27.5	4.0, 23.5	5.3, 14.8	5.3, 9.8
	1	–, 20.4	−1.2, 18.3	4.3, 13.7	4.9, 9.5
3	0	–, 13.1	−5.9, 13.6	2.6, 12.1	3.5, 8.0
	1	–, 14.0	−6.1, 13.6	−0.1, 9.4	3.4, 7.9

Table 5.4: Performance of the PSHF versus target values of jitter σ_T, shimmer σ_A and HNR σ_N; the results are η_v, η_u in dB.

and σ_A, tend to reduce performance. Yet, η_v is positive in all but a few extreme cases.

In summary, the introduction of any form of disturbance, from noise, jitter or shimmer, drastically reduced the performance. Thus, for positive HNR values, the algorithm enhanced the anharmonic component (i.e., improved its SER) much more than the harmonic one, which therefore aids us in the study of unvoiced sound production mechanisms.

5.8 Conclusion

An analysis technique has been developed for decomposing mixed-source speech signals that is based on a pitch-scaled, least-squares separation in the frequency domain. The PSHF technique provides estimates of the voiced and unvoiced components, as harmonic and anharmonic parts, using only the speech signal. The components can subsequently be subjected to any standard analysis, either as time series or as power spectra. Therefore, decomposition is useful because it enables us to analyse the voiced and unvoiced components separately. There are various methods for separating the components, which have been discussed here, but ours is novel because its analysis frame is scaled to a whole number of pitch periods, which gives it performance advantages. It also overcomes the problem of the disparate demands of time series and power spectral analyses by explicitly providing two pairs of output signals.

The PSHF performance was evaluated with three kinds of perturbation: jitter, shimmer and additive noise using different values of $\bar{f}_0$. Tests on synthetic speech demonstrated the PSHF's ability to reconstruct the components, despite corruption by jitter, shimmer and additive noise. The tests across pitch values showed that the performance at otherwise matching conditions was unaffected. The PSHF achieved improvements of $\eta_v = 5\,\mathrm{dB}$ and $\eta_u = 15\,\mathrm{dB}$ to the SER of the harmonic and anharmonic part for parameters typical of normal speech ($\sigma_T = 0.5\,\%$, $\sigma_A = 0\,\mathrm{dB}$ and $\sigma_N = 10\,\mathrm{dB}$), which decreased with increased corruption. For the range of values chosen, fluctuations in the pitch period (jitter) tended to have a larger effect on performance than amplitude fluctuations (shimmer). For positive HNR values, the algorithm enhances the anharmonic component more than the harmonic one, which is of particular interest in the study of unvoiced sound production mechanisms. Evaluation predicts that the harmonic output signals are approximately twice as good as the original signal, in an MSE sense; the anharmonic ones are typically improved by c. 4 dB more than the HNR. For recordings of normal speech, the results suggest improvements to the SER of about a factor of 5 ($\eta_u \approx 14\,\mathrm{dB}$) in the anharmonic component (typically $\sigma_N \approx 15\,\mathrm{dB}$ for vowels, $\sigma_N \approx 3\,\mathrm{dB}$ for voiced fricatives), and $\eta_v \approx 4\,\mathrm{dB}$ for the harmonic component. In the next chapter, we will demonstrate the value of decomposition when applied to the analysis of real speech signals.

Chapter 6

Mixed-source decomposition: Results

6.1 Introduction

This chapter demonstrates the capability of the PSHF to decompose real speech into components that approximate the contributions of the voiced and unvoiced source to the acoustic signal. The algorithm was applied to a variety of sounds which included fricatives and vowels, as well as variations in the mode of phonation, such as pressed and breathy voicing. It represents an exploratory study into the effects of the PSHF and its ability to help extraction of latent features from the speech signal.

Time series were inspected for anomalies and signal features, and short-time (windowed) spectra were computed at points of interest. Spectrograms were also used as a way to identify features in the recorded and processed signals. Ensemble averages were generated by marking equivalent locations in an array of tokens (e.g., at the centre of a fricative or the release of a stop, as before), summing the sound power of the discrete Fourier transform (DFT) of the corresponding windowed portions from each token, and dividing by the number of tokens. For consistency, the first and last token of each group, which tend to exhibit greater variability in stress, emphasis, breath and rhythm, were discarded. Thus consistent features were amplified in relation to others, and also an indication of measurement variability was obtained. Similarly, time averages were generated by averaging the power spectra of consecutive frames, for sustained sounds.

6.2 Recorded speech

To familiarise ourselves with the strengths and weaknesses of the technique, let us begin by examining the result of applying the PSHF to a simple recorded utterance.

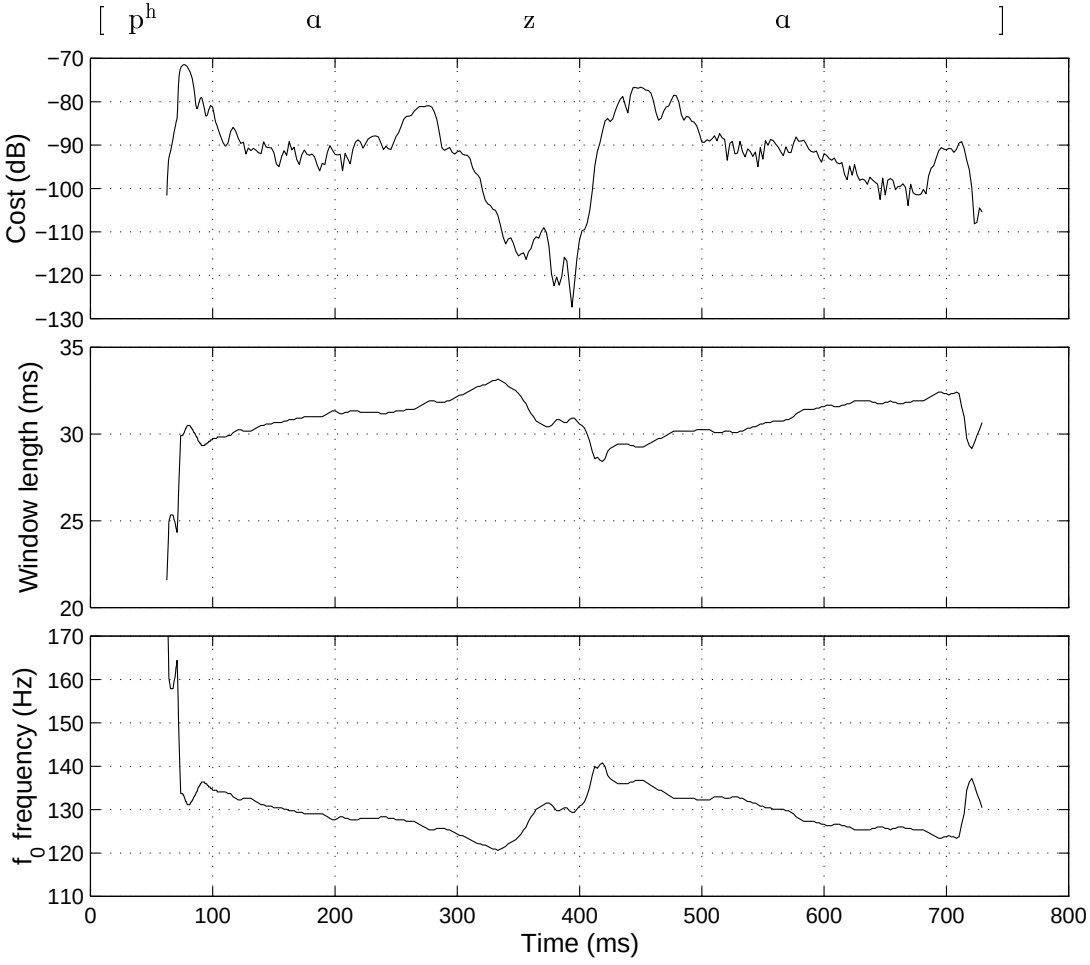


Figure 6.1: Profiles of (top) the minimum cost $J(N(p), p)$, as in Eq. 5.6, (middle) the corresponding window length $N(p)$ expressed in ms, and (bottom) the fundamental frequency $f_0(p)$ during the utterance *C3*-[pʰɑzɑ] by PJ, example #1.

6.2.1 Nonsense word

Our first example, #1, which we use throughout this subsection, is of the nonsense word [pʰɑzɑ] produced by an adult male subject (PJ) for *C3*. Many repetitions of *C3*-[pʰɑzɑ] were processed by the PSHF, of which #1 is typical. The fundamental frequency f_0, plotted in Figure 6.1 (bottom), follows a standard declination during the two vowels, a less stable period during the fricative (320–420 ms) and discontinuities at voice onset and offset, as expected. Above it (Fig. 6.1, middle), the window length, which is four times the pitch period, exhibits reciprocal behaviour, being related by the expression:

$$N(p) = \frac{b f_s}{f_0(p)}, \tag{6.1}$$

where $b = 4$. The minimum cost, which is the value of $J(N(p), p)$ when $N(p) = N_{\text{opt}}(p)$ (as defined by Eq. 5.6 in Section 5.3.2), is plotted in dB in Fig. 6.1 (top). Its overall shape is dominated by the signal amplitude, but exhibits local maxima at transitions. Figure 6.2

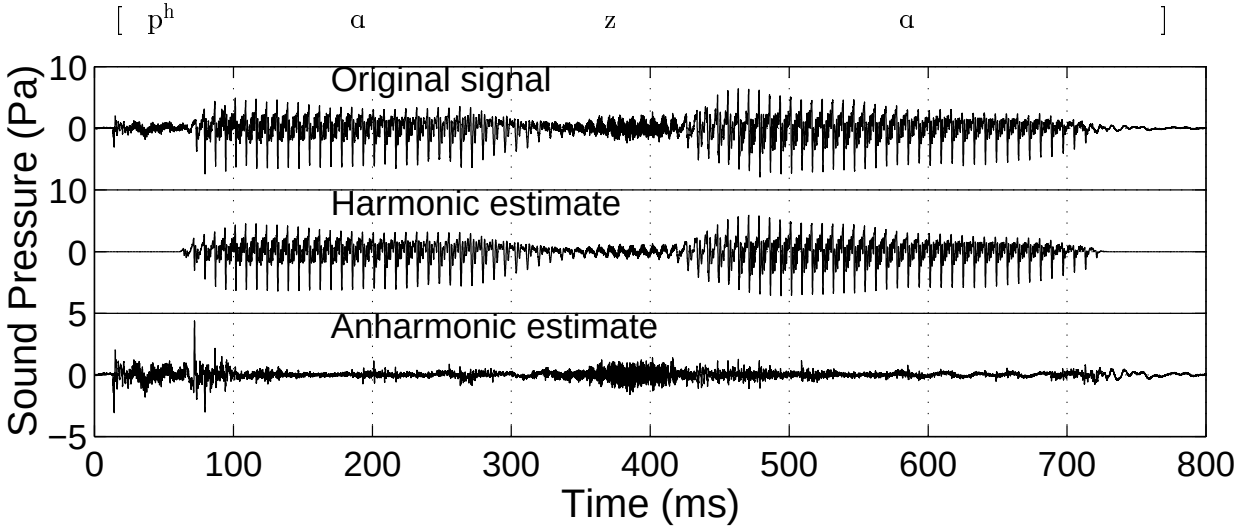


Figure 6.2: Time series of $\mathcal{C}3$-[pʰɑzɑ] from #1 by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom, note amplitude scale) the anharmonic component $\hat{u}(n)$.

contains the time series of the original signal, the harmonic estimate and the anharmonic estimate, respectively; Figure 6.3 shows the spectrograms of the same signals, s, $\hat{v}$ and $\hat{u}$, underneath.

In the voiceless regions (0–70 ms and 720–800 ms), there was no need to extract the voiced component, so the PSHF was not applied. For our purposes the voiced/voiceless decision was made manually, although there are many ways to do so automatically (e.g., Hermes 1988). Therefore, the harmonic outputs were set to zero, $\hat{v} = \tilde{v} = 0$, and the anharmonic outputs were set equal to the original signal, $\hat{u} = \tilde{u} = s$, during the voiceless periods at either end of the utterance.

The original signal of the nonsense word, which is plotted in Figure 6.2 (top), shows the initial burst (at 20 ms) followed by some frication and aspiration leading up to voice onset (at 70 ms), the first vowel (80–320 ms), the voiced fricative (320–420 ms) and the second vowel (420–720 ms). One can see in the harmonic estimate $\hat{v}$ a smooth and clean estimate of the quasi-periodic component, as expected, which has captured the timing of the pulses and tracked gross changes in the envelope. The anharmonic component $\hat{u}$ contains the burst transient and initial noise (20–70 ms), a small amount of noise during the vowels, but the majority of the signal during the fricative, which slowly swells and then dies away. However, there are also glitches, a by-product of processing, at voice onset (70–100 ms) and other transient stages (200 ms, 270 ms, 450 ms), where there are either rapid changes in f_0, amplitude of voicing, or both. Thus, the PSHF algorithm appears to provide the most faithful decomposition during steady spells of voicing, whereas the presence of jitter, shimmer and abrupt changes cause perturbation errors.

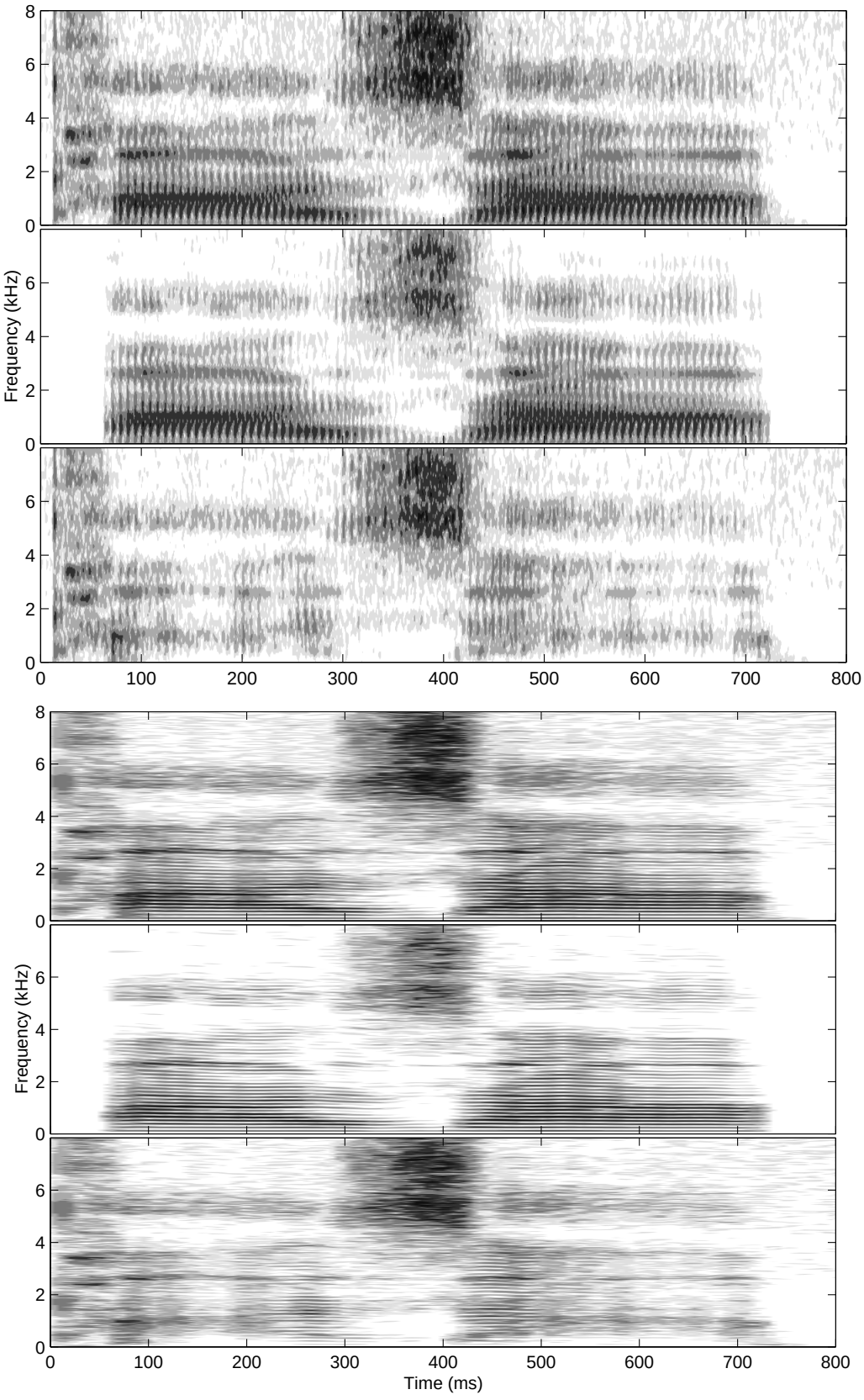


Figure 6.3: Wide-band (upper half, 5 ms) and narrow-band (lower half, 43 ms) spectrograms (Hann window, ×4 zero-padded, fixed grey-scale) of #1 by PJ, *C3*-[pʰɑzɑ], computed (top) from the original signal $s(n)$, (middle) from the harmonic estimates $\hat{v}(n)/\tilde{v}(n)$, and (bottom) from the anharmonic estimates $\hat{u}(n)/\tilde{u}(n)$.

The decomposition is illustrated in Figure 6.3 as two sets of spectrograms of the entire [p$^{\text{h}}$ɑzɑ] utterance. The upper three graphs are wide-band spectrograms of the original signal, harmonic component and anharmonic component, s, $\hat{v}$ and $\hat{u}$, with an effective bandwidth $\Delta f \approx 200\,\text{Hz}$; the lower three are narrow-band spectrograms of s, $\tilde{v}$ and $\tilde{u}$ with $\Delta f \approx 12\,\text{Hz}$.[1] The following acoustic features are visible in the wide-band spectrogram of the original signal (uppermost spectrogram in Fig. 6.3): a vertical stripe (the initial burst) succeeded by broad-band noisy excitation of the formants, the onset of voicing evidenced by further striations at the glottal pulse instants with slowly-varying horizontal bands (at the formants), which continue until the start of the fricative (around 300 ms) where we see voicing dying down to a minimum and the growth of high frequency noise (up to 380 ms), then the second vowel and finally voice offset. Going into the fricative, we see the separation of F1 and F2, as the low back /ɑ/ vowel gives way to the high forward tongue position of /z/, which then became attenuated, consistent with a frication source and weaker voicing. These effects were reversed with the onset of the second /ɑ/. Notice that there also appears to be an anti-resonance whose frequency dips at the end of the first vowel and rises at the start of the second. Its effect can be seen moving from c. 3 kHz at 240 ms at a rate of c. 20 Hz/ms towards a minimum of, perhaps, ~ 1.5 kHz before returning, albeit more slowly, which is most obvious between 440 ms and 540 ms. This trajectory, which is clearly the result of the articulation of the fricative, is probably caused by movements at the back of the tongue in the pharynx and could be attributed to the pyriform sinuses, although we have no further evidence.

The harmonic estimate retains a smaller yet significant part of the frication noise (Fig. 6.3, second from top), but in the vowels the voicing stripes are generally cleaner and more pronounced. The anharmonic wide-band spectrogram (bottom of upper half — third from top) is generally mottled in appearance, which is a characteristic of noisy sounds. However, different frequency regions are excited in each of the four sounds: burst (10 ms, all frequencies instantaneously with lowered formants), aspiration (20–60 ms, all frequencies), mid-vowel (c. 160 ms and 570 ms, principal formants), and frication (320–420 ms, higher formants). It is also possible to see vertical striations in the high-frequency turbulence noise during the onset of frication, which become less noticeable towards mid-fricative. Unfortunately, there is contamination from the voiced part, particularly in unsteady regions (i.e., 200 ms, 270 ms, 450 ms) and at voice onset (70–100 ms), which correspond to rapid changes of f_0 and local peaks in the cost function, as seen in Fig. 6.1.

[1] The signal estimates, $\hat{v}$ and $\hat{u}$, were used to generate the wide-band spectrogram, rather than the power estimates, because the window length (5 ms) does not allow the harmonics to be resolved, negating any benefit from spectral interpolation; the situation is reversed for narrow-band spectrograms, for which $\tilde{v}$ and $\tilde{u}$ were duly used. To decide which pair of outputs to use, we recommend placing a threshold on the DFT frame size at half the mean PSHF window length, $\frac{\langle N \rangle}{2} \approx 15\,\text{ms}$ in this example.

There is a significant contribution to each of the three signals, s, $\hat{u}$ and $\hat{v}$ from a band spanning the 5–6 kHz region during the utterance, which probably encompasses formants F6 and F7. This bar of sound appears to be a mixture of voicing and aspiration sources during the vowels, but was most probably frication in the fricative. In fact, the pulsing within that band, which was in-phase with the glottal pulses during the vowels, underwent a phase shift at the transition into frication at 280 ms. (This is most clearly seen in the harmonic wide-band spectrogram.) Nevertheless, the majority of the structure above 1 kHz in the developed fricative (350–420 ms) was captured by the anharmonic component. Note that, without the aid of any pre-processing or heuristic filtering, the majority of the high-frequency turbulence noise has been passed to the anharmonic component, while the low-frequency voiced part has been successfully allocated to the harmonic component.

Turning now to the narrow-band spectrograms (Fig. 6.3, lower half), one can see the horizontal striations from the harmonics of the fundamental frequency, both in that of the original signal (top) and more obviously in that of the harmonic component (middle). Some of the effects of prosody are visible from these striations, such as when the harmonics cross a formant resonance (e.g., F3 at 2.7 kHz, 100–200ms). As before, the harmonic spectrogram is cleaner than the original, while the anharmonic one (bottom) retains a mottled appearance. Short sections of horizontal striping are evident in parts of the anharmonic (narrow-band) spectrogram, where voicing perturbations have caused some leakage. However, the overall structure of $\tilde{u}$ was not generally periodic, and hence the stripes are absent from the pulsed frication noise, whose envelope alone was periodic. Similarly, throughout much of the vowel sections, the anharmonic spectrogram is not striated (e.g., 590–680 ms in Fig. 6.3, sixth from top), while the corresponding wide-band spectrogram (third from top) shows clear signs of modulation. Again, this implies that the PSHF has extracted pulsed noise into the anharmonic estimate, which would most likely be from aspiration in the case of vowels.

Some features manifest themselves more distinctly in the narrow-band spectrograms, such as the first two formants in the anharmonic component at voice onset (i.e., F1, F2 $\approx$ 1.0 kHz, 1.4 kHz at 80 ms), and the change in formant frequencies from the preceding aspiration (F1, F3 $\approx$ 0.8 kHz, 2.4 kHz at 30 ms; 1.0 kHz, 2.6 kHz at 80 ms). Also, while the lower harmonics of f_0 are virtually eliminated from the anharmonic spectrogram (e.g., 100–250 ms and 470–700 ms), they are shown in the harmonic one, even continuing throughout the fricative (i.e., f_0, $2f_0$ and $3f_0$, 320–420 ms).

Returning to the vowel-fricative transition [-ɑz-], Figure 6.4 gives an expanded view of the reconstructed signals showing the growth of the anharmonic component while the voicing dies down. Compared with the original signal, the harmonic component $\hat{v}$ is much cleaner in

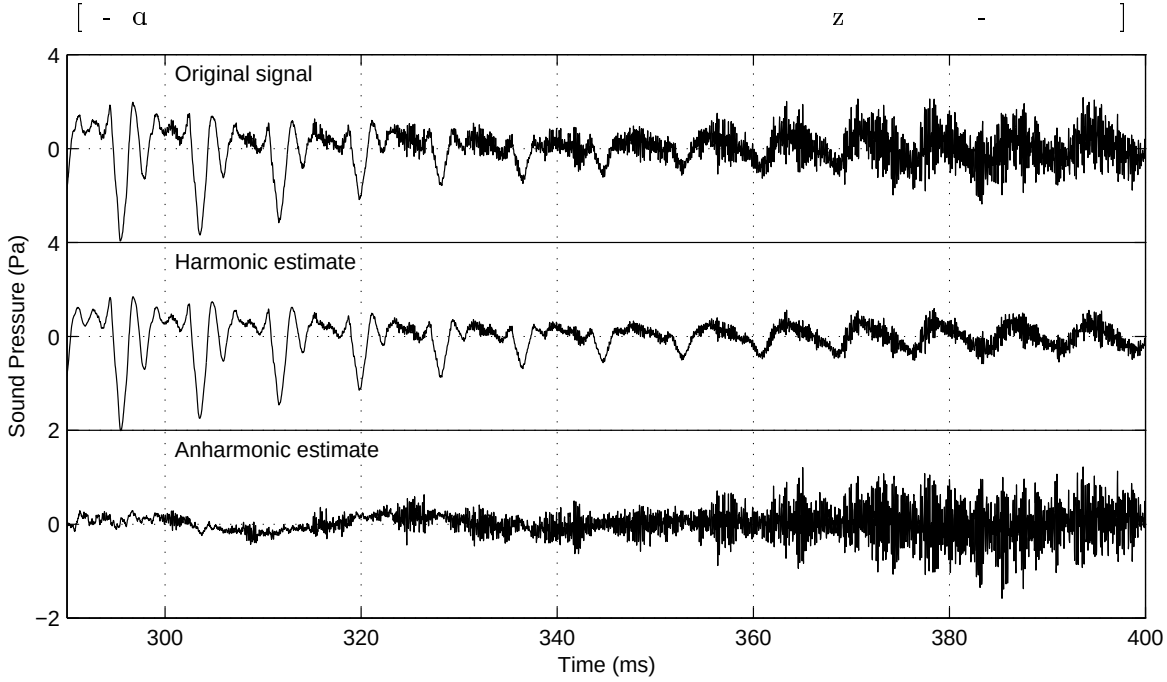


Figure 6.4: Expanded view of time series of $\mathcal{C}\mathit{3}$-[p$^{\text{h}}$ɑzɑ] by PJ from the previous figure, showing the [-ɑz-] transition of the developing fricative, with its PSHF decomposition: (top) original $s(n)$, (middle) harmonic component $\hat{v}(n)$ and (bottom) anharmonic component $\hat{u}(n)$ at double amplitude scale.

appearance, but suffers increasing contamination from frication noise.[2] The regularity of the continuing vocal fold oscillation is obvious, even in the middle of the fricative (c. 380 ms). The periodic pulses in $\hat{v}$ become less spiky, consistent with a weaker glottal closure, and approach the form of a simple harmonic oscillation (that is increasingly contaminated). Although devoicing sometimes occurs in voiced fricatives, it is clear that that is not the case here, since oscillation persists throughout the [z].

The anharmonic component, which is plotted with double the amplitude scale, is very small at the end of the vowel, commensurate with a typically high HNR for modal voice (+17 dB). The HNR drops dramatically (by 20 dB) to about −3 dB, as $\hat{u}$ grows during the transition. In agreement with earlier observations (Fant 1960; Flanagan 1972; Klatt and Klatt 1990; Laroche et al. 1993; Stevens and Hanson 1995), the anharmonic part $\hat{u}$ exhibits modulation by the voice source during development of the fricative (300–370 ms). The effect becomes negligible (around 380 ms) as voicing dies away and the noise level increases; the noise initially comes in bursts with each glottal pulse, then blurs into continuous noise in the fully-developed fricative.

[2] It is possible to incorporate empirical knowledge of speech signals to reduce the cross-contamination of the voiced component, e.g., by low-pass filtering (Laroche et al. 1993), but subjective assessment indicates that additional processing often incurs a loss of intelligibility (Lim et al. 1978).

6.2.2 Summary

For the nonsense word ℂ𝕴-[pʰɑzɑ] by PJ, we have examined time series and spectrograms of the decomposed signals which were used to extract features of the individual components.[3] Examination of the time series at the vowel-fricative transition revealed the weakening of modulation of the anharmonic part as the fricative developed. In sustained fricatives generated by the same subject, however, the modulation persisted. Subjective assessment, from informal listening tests of the separated components, reveals that the harmonic component of [pʰɑzɑ] sounds like [ɑzɑ] with less emphasis on the fricative. The anharmonic component approaches a whispered version of the original [pʰɑzɑ], albeit with some remnants of voicing.

Thus, the PSHF has provided separate output signals that can be analysed individually for feature extraction (d'Alessandro 1990; Richard and d'Alessandro 1997), or in tandem to investigate interactions of voicing and noise sources. Alternatively, the anharmonic component of voiced phonemes can be compared with their voiceless correlates to evaluate differences in their production, as we shall see in the next section.

6.3 Fricatives

Figure 6.5 shows ensemble-averaged spectra of [z] in [pʰɑzɑ] context, its harmonic and anharmonic estimates, and of [s] in [pʰɑsɑ]. It is known that the vocal-tract configuration is very similar for the voiced-unvoiced minimal pair /s, z/ (e.g., Narayanan et al. 1995). So we would expect the spectrum of the anharmonic component of [z] to be similar in shape and amplitude to that of the corresponding unvoiced fricative, [s]. In fact, peaks in the unvoiced fricative [s] spectrum at 1.0 kHz, 1.4 kHz, 1.8 kHz and 2.6 kHz occur in the anharmonic [z] spectrum at a greater amplitude than they occur in the harmonic spectrum, and it is clear that the majority of the energy from voicing, in the range 100–800 Hz has been correctly attributed to the harmonic part with an HNR of ~ 20 dB. Again, the effect of the PSHF confirms what would be anticipated.

6.4 Vowels

This section shows examples of vowels decomposed by the PSHF. All the illustrations are for the vowel [ɑ], although the vowels [i] and [u] were also studied. There were many similarities between the different vowels, but some of the differences are discussed in Section 6.7.

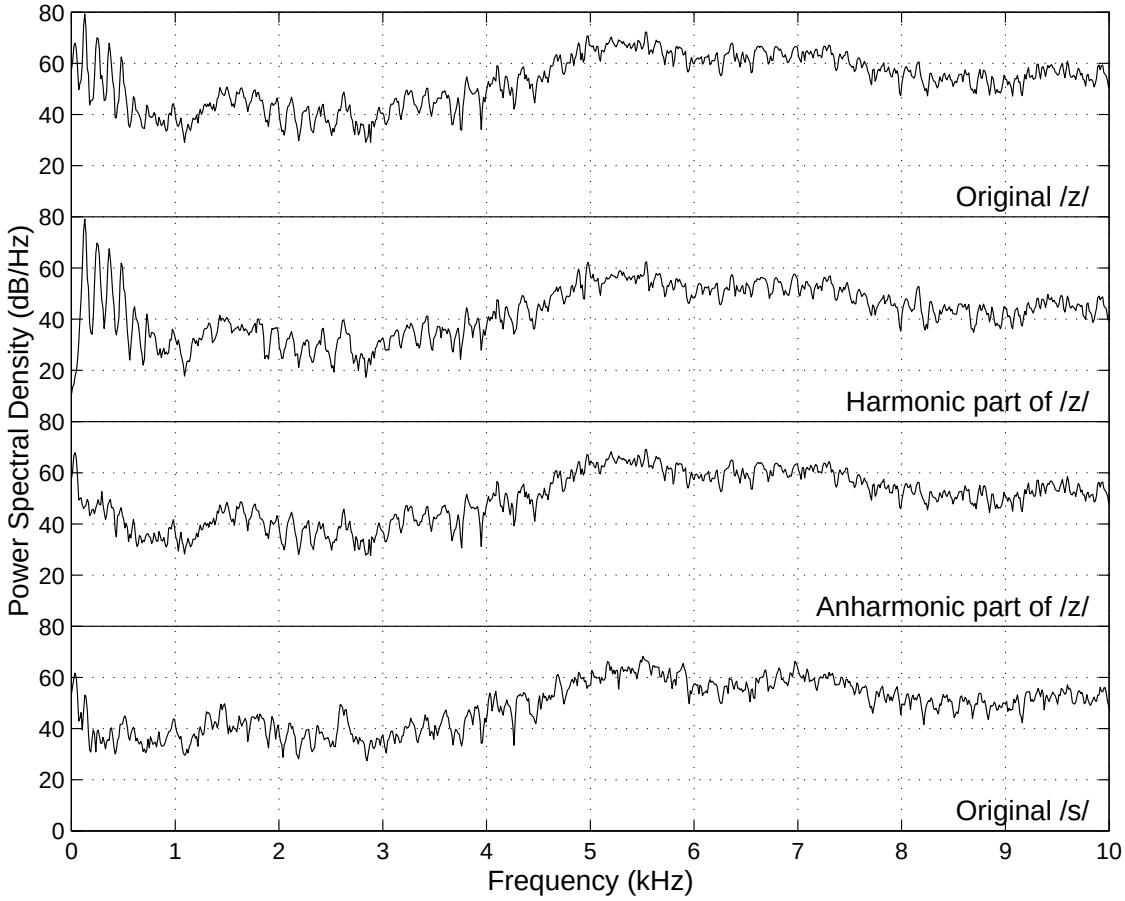


Figure 6.5: Ensemble-averaged spectra of mid-fricative in modally-phonated $\mathcal{C}\mathit{3}$-[p$^{\text{h}}$ɑFɑ] context by PJ (F=/z, s/, 8 tokens, 85 ms window), with the PSHF harmonic and anharmonic decomposition of [z].

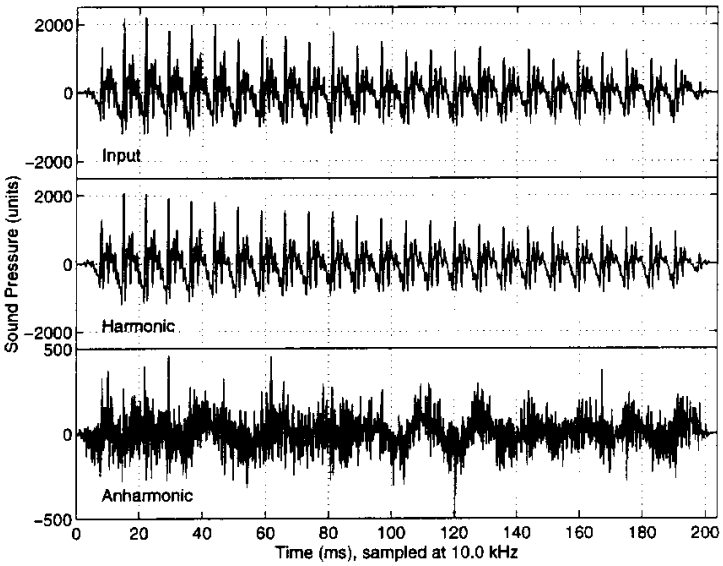


Figure 6.6: Time series of a modal [ɑ] vowel in the context $\mathcal{C}\mathit{1}$-[p$^{\text{h}}$ɑ] by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom, five times amplitude scale) the anharmonic component $\hat{u}(n)$.

6.4.1 Preliminary recordings

Figure 6.6 shows the speech signal, $s(n)$, from a segment of the vowel in the utterance *C1*-[p$^{\text{h}}$ɑ] (by subject PJ) that was processed using the PSHF, in preliminary trials. It also shows the results of that processing: the harmonic component $\hat{v}$, and the anharmonic component $\hat{u}$. On inspection, one can see that the objective of apportioning the quasi-periodic part to the harmonic component and the remainder to the anharmonic, seems to have been broadly met. The harmonic signal $\hat{v}$ is reasonably steady, changing slowly in pitch and amplitude through the recording; the anharmonic signal $\hat{u}$ is noisy and more variable, yet only about a quarter of the amplitude. The shape of the pulses remains fairly constant throughout the segment, suggesting that the vocal-tract filter characteristic is static at this stage in the utterance. There is a low-frequency oscillation in the anharmonic trace, which is most pronounced around 120 ms.

The anharmonic signal in Figure 6.6 contains periodic noise at low frequency, in the 40–50 Hz region. A first guess would suggest that it is interference from the mains electrical supply (50 Hz), but a closer analysis revealed that there was more than one periodic constituent. In fact, the frequency of the second lesser spike ~ 44 Hz matches the oscillation visible in the $\hat{u}$ time series at 120 ms. This phenomenon was investigated indicating that the noise was probably generated by air flow impinging on the microphone, causing wind noise. Care was taken in later recordings (*C3* onwards) to position the microphone either at an angle away from the air stream expelled by the subject or far enough from the lips that the effect was negligible.

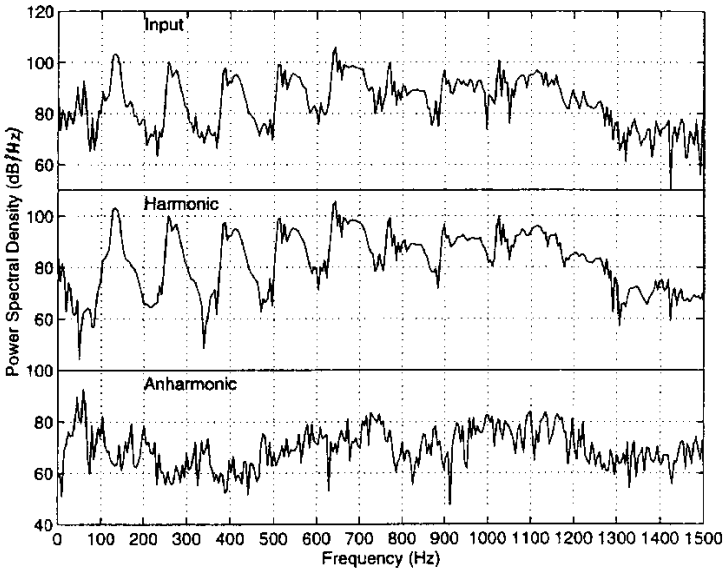


Figure 6.7: Power spectra of a modal [ɑ] vowel in the context *C1*-[p$^{\text{h}}$ɑ] by PJ: (top) the original $S(k)$, (middle) the harmonic component $\hat{V}(k)$, and (bottom) the anharmonic component $\hat{U}(k)$.

The spectra of these signals, S, $\hat{V}$ and $\hat{U}$, were calculated for the entire vowel segment

[3] Sound files can be found at the project web site (Jackson 1998).

(213 ms duration), and are shown in Figure 6.7 up to 1.5 kHz. The breadth of the harmonic peaks at multiples of the fundamental frequency ($f_0 \approx 130$ Hz), as seen in S and $\hat{V}$, was the result of the pitch variation from approximately 140 Hz to 125 Hz, whose effect increases for the higher harmonics. It is reassuring to note that by far the majority of the sound power in the range 100–1250 Hz corresponds to the harmonic component, as we would expect. The envelope of the input spectrum $S(k)$ and indeed the other spectra, $\hat{V}(k)$ and $\hat{U}(k)$, exhibit formant resonances at F1 $\approx$ 720 Hz, F2 $\approx$ 1.1 kHz, etc. Above 1.5 kHz, the ambient noise level and the roll-off of the signal leave little useful information in the spectra from these recordings. The anharmonic spectrum (Fig. 6.7, bottom) appears to undulate with a gross ripple most evidenced by the spectral humps near 200 Hz, 340 Hz and 480 Hz. Comparing it to the harmonic spectrum (middle), we see that the peaks of one correspond to the troughs of the other, and vice-versa. This is an artefact of the anharmonic estimate's spectrum $\hat{U}(k)$, and illustrates the need for interpolation, when performing narrow-band spectral analysis.

Another point to note about the decomposition of these preliminary recordings (i.e., *C1* and *C2*) is that the zeroth harmonic was assigned to the harmonic part; whereas, as the algorithm is defined in Chapter 5 (and as it has been applied to all subsequent recordings), the first DFT bin, which corresponds to the d.c. component, is given to the anharmonic part. Hence, for frequencies $f < f_0/4$, the majority of the signal is assigned to $\hat{v}$, as seen in Figure 6.7, although it passes largely unnoticed in the harmonic signal. The anharmonic signal $\hat{u}$, on the other hand, is dominated by the residual low-frequency wind noise $f_0/4 \leq f \leq 3f_0/4$, in this case, plainly evident in Figure 6.6.

6.4.2 Sustained vowel

Our second example, #2, a sustained vowel *C4*-[ɑ] produced by an adult female subject, SB, was decomposed to give the harmonic and anharmonic estimates, $\hat{v}$ and $\hat{u}$, and the power-based estimates, $\tilde{v}$ and $\tilde{u}$ respectively. Figure 6.8 depicts the spectra derived from the original signal s, and the latter output pair, $\tilde{v}$ and $\tilde{u}$, using a steady section from the centre of the vowel. Note that this time almost all of the very low-frequency power has been attributed to the anharmonic component, as it should be.

The periodicity of $\tilde{v}$ is strongly marked by the harmonic peaks of its spectrum, still noticeable above 8 kHz. Reassuringly, the levels of the harmonic peaks remain practically untouched by the PSHF, while the inter-harmonic troughs were deepened. Both components show the effect of the principal formants, although their spectral tilts are very different. Apart from the very low-frequency noise ($f < 50$ Hz, mostly wind noise generated at the microphone), $\tilde{u}$ contains a much greater portion of the original signal at high frequencies ($f > 3$ kHz), as expected for flow-induced, turbulence noise. Moreover, in the detail, there are features distinct to the

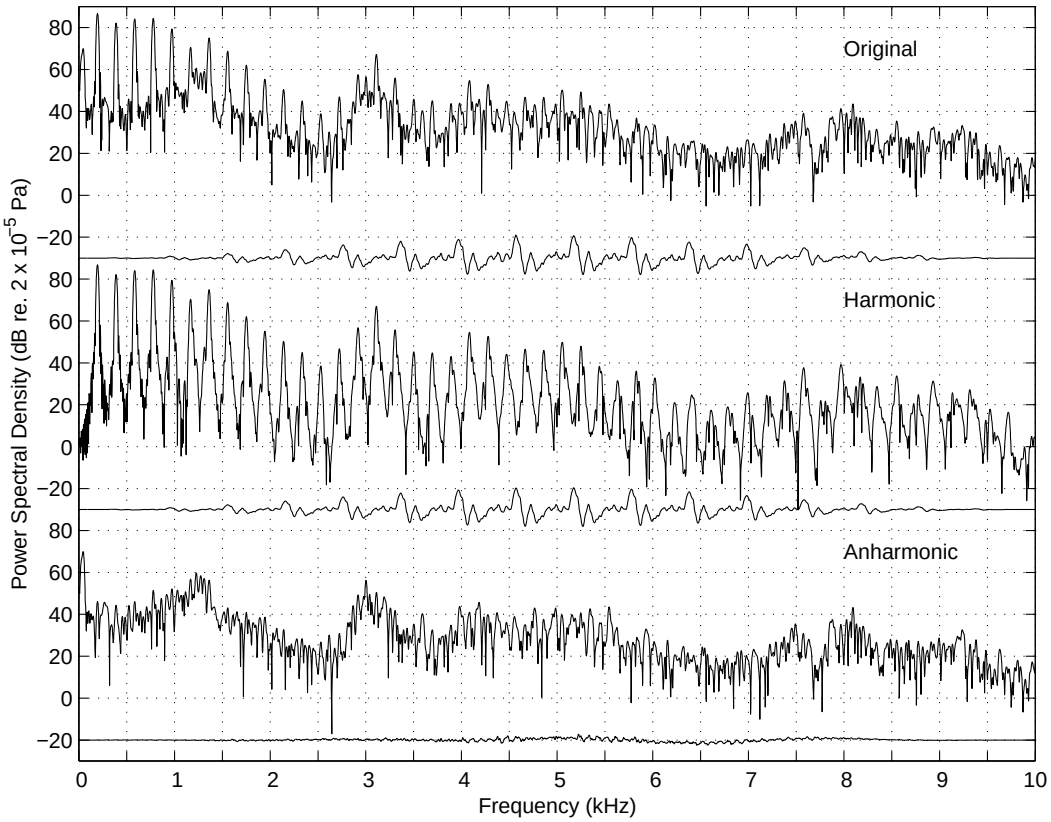


Figure 6.8: Power spectra (85 ms, Hann window, $\times 4$ zero-padded) computed from the original signal $s(n)$ (top) from the modal vowel [ɑ] by female subject SB, the harmonic estimate $\tilde{v}(n)$ (middle) and the anharmonic estimate $\tilde{u}(n)$ (bottom), whose time series are inset underneath each graph (anharmonic signal drawn at double amplitude).

anharmonic spectrum, such as a peak which had been hidden between the first two harmonics ($\sim 250\,\mathrm{Hz}$) and a trough just above F2 at $1.4\,\mathrm{kHz}$.

The jitter, shimmer and HNR were measured locally for the same section of speech: $\tilde{\sigma}_T = 0.9\,\%$, $\tilde{\sigma}_A = 0.07\,\mathrm{dB}$ and $\tilde{\sigma}_N = 14\,\mathrm{dB}$. The values of jitter and shimmer are typical for a normal healthy voice in sustained phonation (Blomgren et al. 1998), as is the HNR, although is somewhat lower than we measured for the male voice. These values were used to predict the PSHF's performance by interpolating the results of Table 5.4 and Figure 5.20, to give $\eta_v \approx 3\,\mathrm{dB}$ and $\eta_u \approx 17\,\mathrm{dB}$. Thus, we can claim with some confidence that we have improved the estimate of the voiced part over the original signal having halved the noise power, and that the majority of the unvoiced part was produced by an unvoiced source. In contrast to the spectra of $\tilde{v}$ and $\tilde{u}$, $\tilde{V}(k)$ and $\tilde{U}(k)$ do not reflect the periodic ripples of the harmonics, since the process of spectral interpolation in the second stage of the PSHF algorithm compensates for the gaps that would have otherwise been in the harmonic bins (as in Fig. 6.7).

Both the performance predictions and the interpretations of the harmonic and anharmonic spectra (Fig. 6.8) present a compelling argument for their validity, which is supported by a previous study (Jackson and Shadle 1998, results in Fig. 6.5) that showed good agreement between the anharmonic component of a voiced fricative [z:] and the corresponding unvoiced fricative [s:] produced by the same subject. Examples of other phonation modes for [$\mathrm{p^h}$ɑsɑ] were examined (e.g., breathy, pressed), and showed similar but exaggerated features, since the relative magnitude of the anharmonic components was greater, as was the degree of jitter and shimmer.

6.5 Mode of phonation

To demonstrate the effect of the phonation mode, the PSHF was used to process a modal and a pressed realisation of the vowel in the syllable context [$\mathrm{p^h}$ɑ] by PJ, from the second corpus, *C2*. Whispered speech was not processed, since there was no voicing and thus no voiced component.

6.5.1 Modal

Figure 6.9 shows the segment from a typical modally-spoken token, which was processed by the preliminary version of the PSHF. As before, the separation of the harmonic part from the input signal is quite effective, despite the sizeable change in f_0 during the segment (from about 185 Hz to 115 Hz, cf. 140 Hz to 125 Hz in Fig. 6.6). However, some of the low-frequency oscillations, which are undeniably still present, appear in both output components, and there are some glitches in the anharmonic component that are clearly related to the glottal pulses (e.g., at 470 ms, 478 ms and 487 ms). Apart from the much larger low-frequency transient

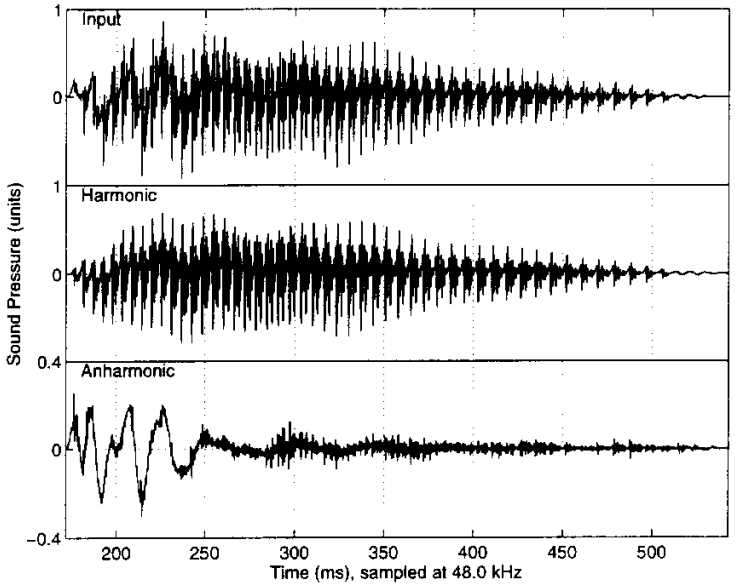


Figure 6.9: Time series of a modal [ɑ] vowel in the context *C2*-[p$^{\text{h}}$ɑ] by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom, amplified scale) the anharmonic component $\hat{u}(n)$.

after voice onset, the ratio of the harmonic to anharmonic signal amplitudes is about 10:1, i.e., HNR ≈ 20 dB, similar to the earlier recordings.

The low-frequency oscillations do not seem to behave like a narrow-band process, tuned to a specific resonance frequency, although the strongest oscillation (after 200 ms at 50 Hz) lasts around 20 ms corresponding with the earlier observations. Rather, the response is as might be expected from an impulse that had travelled through a dispersive medium. The higher group velocity of higher frequencies is typical of such propagation (e.g., surface waves on the sea or flexural waves in a steel beam) and a feature of the non-acoustic convection of pressure waves in air. In fact, if the impulse occurred at release of the stop consonant [p], then the oscillations arrived about 60 ms later, corresponding to a net velocity of 5 m/s, while the acoustic wave at 340 m/s would have taken 1 ms to reach the microphone.

To give a broader impression of the spectral shaping caused by the vocal tract, rather than the detail of the individual harmonics of the varying f_0, the power spectra were computed from a shorter frame of the relevant signals (85 ms). The window location was chosen to be the most stationary part of the time series (from 270 ms to 355 ms), by inspection. The spectra, plotted in Figure 6.10, exhibit most of the same salient features as those obtained from the *C1* recordings: large harmonic spikes in S and $\hat{V}$, broadening with increasing harmonic number, and formant peaks at F1 $\approx$ 700 Hz, F2 $\approx$ 1.1 kHz, and at F3 $\approx$ 2.7 kHz, and possibly F4 $\approx$ 3.5 kHz and F5 $\approx$ 4.5 kHz. There is also a trough just above 4 kHz, and perhaps another at 4.9 kHz, which could correspond to anti-resonances of the tract. It has been suggested that anti-resonances in

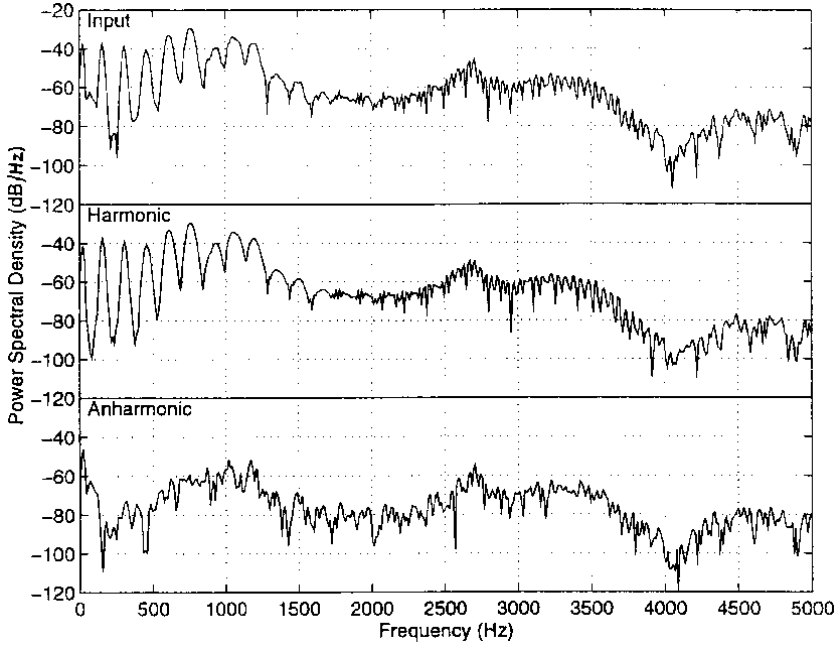


Figure 6.10: Power spectra of a modal [ɑ] vowel in the context *C2*-[pʰɑ] by PJ, using a 4096-point Hann window (≈ 85 ms), centred at 312.5 ms: (top) original $S(k)$, (middle) harmonic component $\hat{V}(k)$, and (bottom) anharmonic component $\hat{U}(k)$.

this range are attributable to the effect of the pyrifom sinuses (Mermelstein 1967; Dang and Honda 1997). The anharmonic spectrum $\hat{U}$ does not appear to ripple, as before, in opposition to the peaks in the harmonic spectrum, except below 500 Hz. Thus, the effects of this artefact are less evident in wider-band analyses, although still present here despite the blurring caused by pitch declination. The low-frequency spike (shown in less detail this time) occurs in both $\hat{U}$ and $\hat{U}$, but this time at 21 Hz. There is a second hump, which occurs around 50 Hz in the spectrum of the anharmonic trace, $\hat{U}$, which suggests that the convective pressure wave (if the above arguments are accepted) may be exciting certain modes or may have, vortices whose size are governed by some characteristic length or physical dimension of the subject.

6.5.2 Pressed

The pressed speech recordings can be described as intense, hoarse speech, somewhat akin to a stage whisper. They sounded more like the speaker (PJ) was desperate for air, rather than the throat had been relaxed to allow a greater air flow rate (as in breathy voice, Abercrombie 1967, for example). A token typical of these recordings was processed using the preliminary version of the PSHF, and the time series are depicted in Figure 6.11.

The pressed speech appears much more noisy to begin with than any of the modal samples, and this is reflected in the overall ratio of the harmonic and anharmonic signals, which is generally much less (HNR $\approx$ 10 dB). The low-frequency 'puff', whose main oscillation has

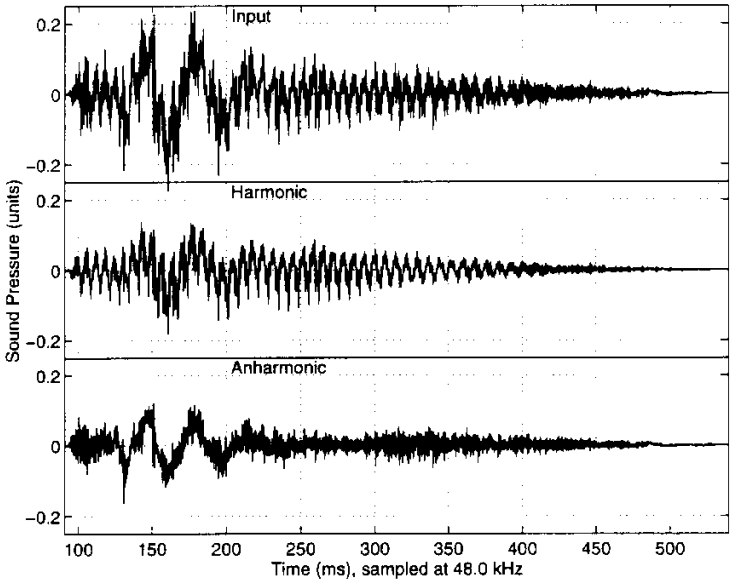


Figure 6.11: Time series of a pressed [ɑ] vowel in the context *C2*-[p$^{\text{h}}$ɑ] by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom) the anharmonic component $\hat{u}(n)$.

a period of $\sim 37\,\text{ms}$ (corresponding to $27\,\text{Hz}$), is present in almost equal proportions in the processed signals. Again there is a second hump in the anharmonic spectrum, which occurs at $\sim 80\,\text{Hz}$. The harmonic component is much quieter than the modal case, while the noisy part of the anharmonic component is much the same as before. The glottal-pulse-related glitches, which are an artefact from inaccurate pulse prediction (caused by perturbations), are apparent between 220 ms and 280 ms, but much less so elsewhere in the record, where the effect is probably masked by the noisy signal. The envelope of the anharmonic component is shaped during the course of the utterance, rising to maxima at the beginning ($\sim 110\,\text{ms}$) and as phonation fades away ($\sim 330\,\text{ms}$), but although it decays as voicing dies away, there is a stage when the anharmonic signal is dominant (400–470 ms). The minimum c. 275 ms coincides with the most stable part of the sample during voicing, when the signal was most like the modal case, although quieter.

The spectra in Figure 6.12 of both the original signal and the output, harmonic component, S and $\hat{V}$ are less regular than the even harmonic peaks of the corresponding modal spectra (Fig. 6.10). They are also less smooth to the point that it is hard to find a harmonic peak in the original spectrum above the sixth harmonic (i.e., $f > 1\,\text{kHz}$). The higher formants remain reasonably well-defined in all three spectra: F3 $\approx 2.7\,\text{kHz}$, F4 $\approx 3.6\,\text{kHz}$, and F5 $\approx 4.8\,\text{kHz}$. The first two formants, on the other hand, were estimated by inspection as F1 $\approx 500\,\text{Hz}$ and F2 $\approx 900\,\text{Hz}$ with more difficulty (cf. 700 Hz and 1.1 kHz respectively, for modal phonation). Thus it appears that, by altering the mode of phonation, the formants may have shifted. LPC

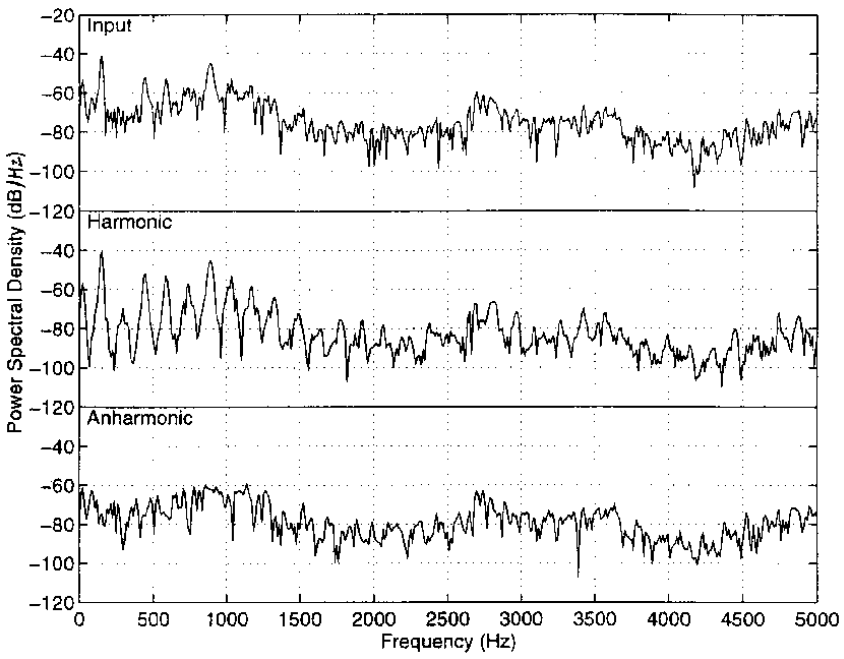


Figure 6.12: Power spectra of a pressed [ɑ] vowel in the context *C2*-[pʰɑ] by PJ, using a 4096-point Hann window ($\approx$ 85 ms), centred at 260 ms: (top) original $S(k)$, (middle) harmonic component $\hat{V}(k)$, and (bottom) anharmonic component $\hat{U}(k)$.

analysis can be useful for picking out the formants and estimating their bandwidths. The sharp trough at 4.1 kHz, that was so clearly present in the modal case, is not obvious in the pressed case, but a broad shallow trough does occur near 4.1 kHz. It is possible that there are noise sources from other locations masking the zero, or that its frequency has changed slightly over the course of the utterance and the shallowness is caused by averaging. Another explanation could be that the losses, and hence bandwidth, increase in the pressed mode.

6.6 Voice quality in vowels

Using the complete version of the PSHF on the later recordings, we compared and contrasted many utterances of the form /$\mathrm{CV_1FV_2}$/, where C was an unvoiced plosive, $\mathrm{V_1 = V_2}$ = /ɑ, i, u/ were vowels, and F a fricative. Figures 6.13 and 6.14 are typical examples of /pɑsɑ/ that were spoken in modal and breathy modes, respectively.

In the modal utterance (Fig. 6.13), the decomposition of [pʰɑsɑ] is very similar to that of the nonsense word [pʰɑzɑ] that we saw earlier (in Fig. 6.2) except, of course, that voicing is totally absent from the fricative here. Although the algorithm should have been disabled during all voiceless regions, this was not done in the fricative to avoid any potential subjective manipulation of the transitions at offset and onset. The PSHF used interpolated pitch estimates when a valid estimate was not available. Normally, all of the input signal would be assigned to

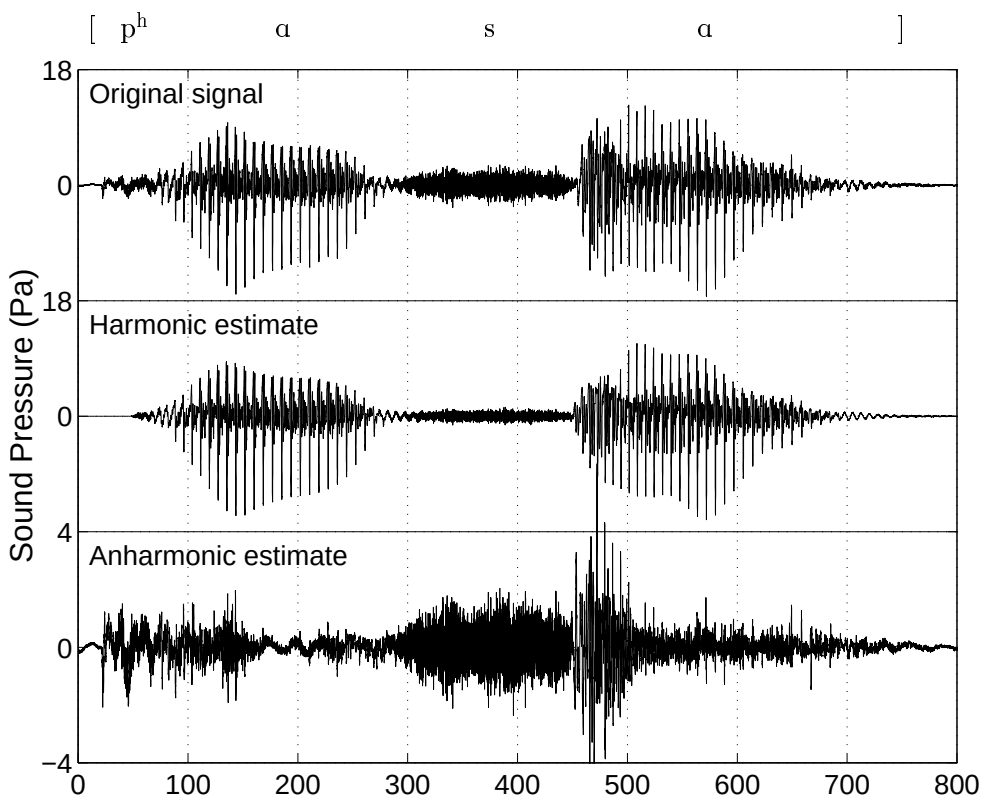


Figure 6.13: Time series of modal $\mathcal{C}\mathcal{Z}$-[pʰɑsɑ] by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom, enlarged scale) the anharmonic component $\hat{u}(n)$.

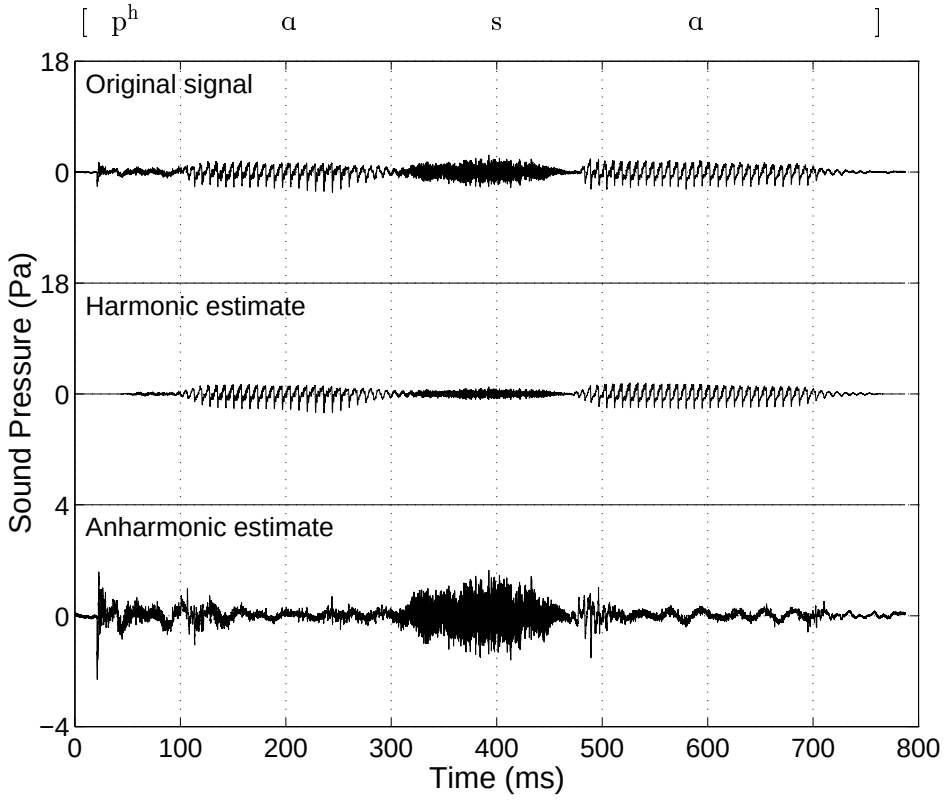


Figure 6.14: Time series of breathy $\mathcal{C}\mathcal{Z}$-[pʰɑsɑ] by PJ: (top) the original signal $s(n)$, (middle) the harmonic component $\hat{v}(n)$, and (bottom, enlarged scale) the anharmonic component $\hat{u}(n)$.

the anharmonic component during periods with no voicing.

Both the timing and the amplitude of the voiced pulses were well-modelled by the harmonic component $\hat{v}$ and, although the envelope was smoothed by the processing, the overall separation is convincing. The anharmonic component $\hat{u}$ was largest as a result of perturbation errors from the second voice onset (460–500 ms), yet other glitches caused by these variations in the pitch and amplitude of voicing were evident throughout much of the voiced portion, in the vowels. Following the onset of the first vowel (c. 70 ms), $\hat{u}$ was a mixture of breath noise and glitches. Later, during steady phonation (150–250 ms), the noise component was still present, yet at a much lower level. The noise grew as the fricative developed and $\hat{u}$ claimed the majority of the fricative signal (300–460 ms).

In comparison, the breathy recording (Fig. 6.14) was generally quieter, and had a much lower HNR: although $\hat{v}$ was greatly reduced compared to the modal case, the effect on $\hat{u}$ was lesser. In fact, the unvoiced sounds (i.e., the consonants) were much less affected by the change of phonation mode to breathy. The perturbation errors, in a similar pattern as before, were significantly reduced by the softer voicing.

In general, breathy utterances, tended to be quieter and to have a lower HNR. In our collection of tokens, the transition into the second vowel was briefer in relation to modal tokens, and there was greater variability in the amplitude of the fricative. This may have been the result of the speaker having finer control over modal realisations. In the pressed tokens, the HNR was lower still, so that most transient errors were masked by the noise in the anharmonic component. The frication noise was slightly louder than in the modal case, and the much quieter vowels had highly variable amplitudes and often double humps in their envelopes.

6.7 Vowel context

This section provides a summary of detailed observations from modal recordings of /pɑsɑ/, /pisi/ and /pusu/. At least eight repetitions of each utterance were used in the comparison, which was designed to allow the PSHF to augment the study of effects of the different contexts provided by the vowels /ɑ, i, u/.

For [pʰɑsɑ], there were brief, small-amplitude transient errors at the onset of the first vowel. The noise component was very low toward mid-vowel, growing substantially to a constant amplitude in [s]. The onset of the second [ɑ] created a large transient with errors persisting until offset. The vowel's envelope comprised a single hump, rising once and then falling.

The initial onset, in [pʰisi], was abrupt and led to slightly larger errors. Those occurring in the vowel after onset did not appear to be related to changes in f_0, and so other changes, such as in the formants, may have been responsible. As before, $|\hat{u}|$ reached a minimum before the

development of [s], which was sustained at approximately constant amplitude. Errors occurred at the second onset of a similar size to those of the first, disappearing in the course of the vowel. The [i] envelopes tended to exhibit double peaks.

The [p^{h}usu] tokens started with a low amplitude transient from voice onset. Breath noise continued throughout the first [u], but the swelling and fully-developed frication were typical. The second [u] transient was very brief and spiky, and was followed by variable noise through the vowel. The envelope of the first vowel tended to have a double hump, while the second usually had a single hump.

Overall, the generall behaviour of the decomposition was consistent for all three vowel contexts. However, differences in the degree and duration of the perturbation errors may provide clues to other aspects of variation, as hinted above, and should be investigated.

6.8 Conclusion

Processing real speech examples resulted in convincing decompositions, extracting and revealing features particular to the individual components. The PSHF produced plausible decompositions of a variety of utterances, including breathy vowels, sustained fricatives and entire nonsense words. The harmonic component retained a noticeable amount of noise, especially in voiced fricatives, whose ensemble-averaged spectrum was similar in shape to (though weaker than) the anharmonic one, at the higher frequencies. In general, the algorithm performed best on sustained sounds; tracking errors at rapid transitions, and errors due to jitter and shimmer, were spuriously attributed to the anharmonic component. Nevertheless, this component clearly revealed various features of the noise source, such as, in voiced fricatives, spectral peaks below 3 kHz (as in Fig. 6.5) and modulation.

Both the performance predictions from Chapter 5 and the intuitive interpretations of the spectra present a compelling argument for the fidelity of the harmonic and anharmonic spectra, which is supported by our results (see Fig. 6.5) that show good agreement between the anharmonic component of a voiced fricative [z:] and the corresponding unvoiced fricative [s:] produced by the same subject. For the nonsense word [p^{h}ɑzɑ] ($^{\#}$1), used spectrograms of the decomposed signals to extract features in the individual components.[4] Examination of the time series at the vowel-fricative transition revealed the weakening of modulation of the anharmonic part as the fricative developed. In sustained fricatives generated by the same subject, however, the modulation persists. Thus, the PSHF method enables investigation of subtle differences in sound production which may shed light on the interaction of voice and noise sources.

The main limitations of the technique concern its computational efficiency, and robustness

[4]Sound files of the examples can be found at the project web site (Jackson 1998).

to deviations of the input speech signal from periodicity. The current implementation is far from real-time, although there is plenty of scope for reducing the amount of computation. Jitter, shimmer, transients and voice onset/offset transitions all tend to degrade performance, although a high degree of robustness has been demonstrated across normal speech conditions. Indeed, local measurements of the perturbation of the original speech signal were used to predict the accuracy of the decomposed signals as estimates of the voiced and unvoiced components. Further work is needed to examine performance enhancements, and to benchmark the PSHF against other methods. However, this chapter demonstrates the potential for applying the PSHF to a variety of speech problems, particularly the analysis of mixed-source speech production and speech modification. Encouraged by these preliminary findings, the following chapter describes a more detailed study into the properties of voiced fricatives, looking specifically at dynamic features revealed by decomposition.

Chapter 7

Mixed-source analysis of fricatives

In Chapter 5, we described an algorithm, the pitch-scaled harmonic filter (PSHF), that decomposes speech into harmonic and anharmonic signals, representing the voiced and unvoiced components respectively. The PSHF was developed from a measure of harmonics-to-noise ratio (HNR, Muta et al. 1988) to provide full reconstruction of harmonic (estimate of voiced) and anharmonic (estimate of unvoiced) time series, on which subsequent analyses can be performed independently. This method is especially suited to acoustic analysis of sustained sounds with regular voicing (i.e., with low values of jitter and shimmer), because of the underlying harmonic model of the voiced part.

Having separated the harmonic and anharmonic components as a means of estimating the voiced and unvoiced contributions to the speech signal, we are well placed to perform individual analyses. Indeed, rather than analysing the components individually or comparing their overall levels, we can begin to examine some of the interactions, which may give us further clues as to the nature of sound production. Voiced fricatives are a good starting point, because phonemes can be produced in an almost stationary configuration, and the principal sources are both relatively well-defined.

The tests with synthetic signals showed that, although the envelope amplitude is slightly reduced with respect to the input signal, its phase remains unaltered. This finding supports our assertion that any modulation exhibited by the anharmonic component is not a processing artefact, but a property of the source component from which it is derived. It should therefore be incorporated into the workings of any noise synthesis procedure, whenever voicing occurs.

In this chapter, we employ the PSHF to study the interaction between sources in voiced fricatives, to arrive at better source models, and to obtain clues to the production mechanism that governs the interaction. Section 7.1 presents preliminary results of the decomposition using data from Corpora 4 and 5 (see Section 4.1 for recording details). Section 7.2 presents further analysis by considering the modulation of the aperiodic component in voiced fricatives, for which results are given in Section 7.3. These results are discussed in light of possible aero-

acoustic mechanisms in Section 7.4. In Section 7.5, we attempt to synthesise a voiced fricative and Section 7.6 concludes.

7.1 Characterising the components

This section illustrates three kinds of analysis that can be employed to extract descriptive parameters from a mixed-source speech signal. Using the example of a sustained fricative, we consider time series, power spectrum and signal power.

7.1.1 Decomposition

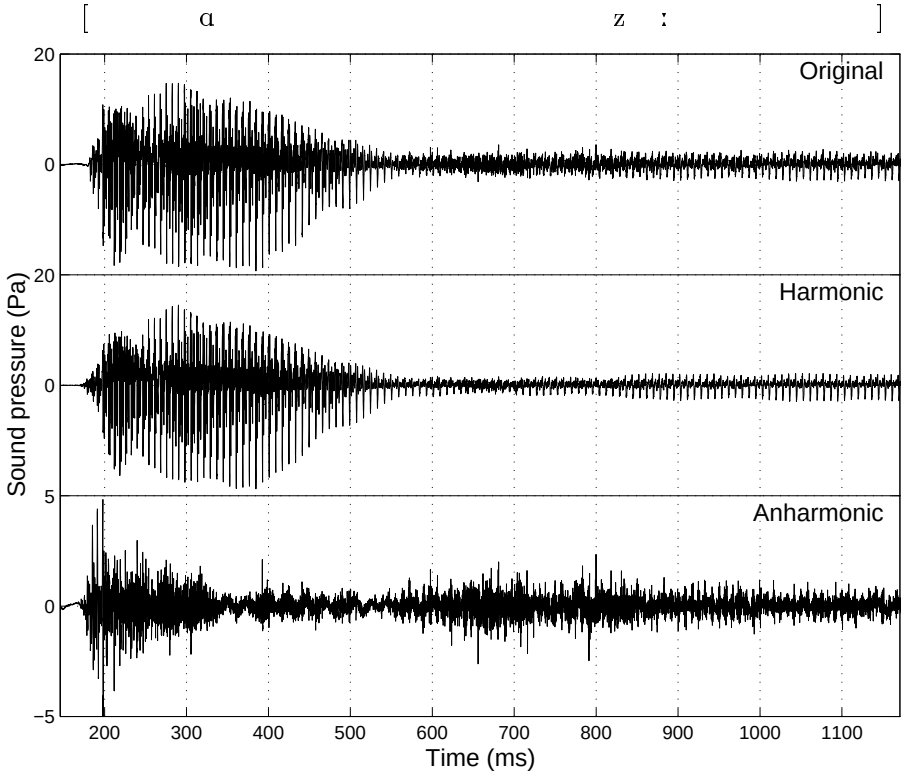


Figure 7.1: Time series from $\mathcal{C}\mathfrak{Z}$-[ɑzː] by an adult male (PJ, #2) of the original signal $s(n)$ (top), the harmonic component $\hat{v}(n)$ (middle), and the anharmonic component $\hat{v}(n)$ (bottom). Note the different amplitude scales.

The vowel-fricative transition [ɑzː] produced by subject PJ was decomposed by the PSHF, as illustrated in Figure 7.1. The majority of the signal energy is modelled by the harmonic component $\hat{v}$, which begins with a rapid growth of voicing that is then sustained at a high level during the vowel. After 200 ms, it starts to fade as the transition is made into the fricative, which appears to achieve a steady state from c. 560 ms onwards. The anharmonic component $\hat{u}$ is of a much lower amplitude in the vowel, although magnified four times in the graph, and follows a very different pattern: with the greatest amplitude initially, it quickly decays to its

minimum in the latter part of the vowel, and reverts to an intermediate magnitude for the fricative, which reduces gradually. The signal $\hat{u}$ is noisy, in contrast to $\hat{v}$ which exhibits a regular pulsing throughout. Each of these general characteristics is as expected, including the initial surge of unvoiced noise, which could be generated by increased airflow at voice onset, although irregularities in phonation would also contribute some spurious elements (as indicated by the tests with synthetic signals).

To extract meaningful information from the component signals, we might first consider their overall amplitudes. By comparing them we obtain an indication of how noisy or periodic the original signal was. Averaged over an entire utterance, this gives us little information about trajectory dynamics or indeed any interaction, but looking at the short-time power for the two signals in parallel can be a way of usefully summarising a particular aspect of the time series that was observed in the previous chapter, namely modulation. We can also consider features of the vocal tract filter, which can be used to identify certain source characteristics such as place. However, the ability to perform parallel analyses of the harmonic and anharmonic components opens new avenues of investigation that give us a new perspective on interactions in mixed-source sound production and may offer a glimpse of their mechanics.

7.1.2 Spectral envelope

A popular means of computing the spectral envelope is linear predictive coding (LPC), which fits an all-pole model to the signal data in a least-squares way, as described in Chapter 4. Now that the signal has been decomposed into two components, separate LPC coefficients can be computed for each part.

Short sections of the signals around 900 ms were used to produce power spectra from the original and the two power-based estimates, $\tilde{v}(n)$ and $\tilde{u}(n)$. The spectra, each overlaid with the results of an LPC analysis (50-pole autocorrelation, to correspond to the 48 kHz sampling rate), are plotted up to 8 kHz in Figure 7.2, with their time waveforms inset. The waveforms show how the harmonic signal $\tilde{v}$ has been purged of noise, which arises in the anharmonic part $\tilde{u}$ as pitch-synchronous packets. Most of the energy in the original spectrum comes in the first five harmonics but, even though the spectrum becomes more noisy at higher frequencies, there is a significant proportion in the range 4–8 kHz. However, the harmonic spectrum maintains its periodic structure over all the frequencies plotted, while the anharmonic spectrum, being pervasively noisy, is devoid of harmonics. Although the smoothed LPC spectra display many similarities, there are notable differences in the resonance frequencies (e.g., peaks differ by 50 Hz at F2, by 200 Hz at F3). Moreover, the first formant F1 is absent from the anharmonic curve, where their relative amplitudes are more than 30 dB apart, which is compatible with the net low-frequency anti-resonance excited by a frication source. At higher frequencies the

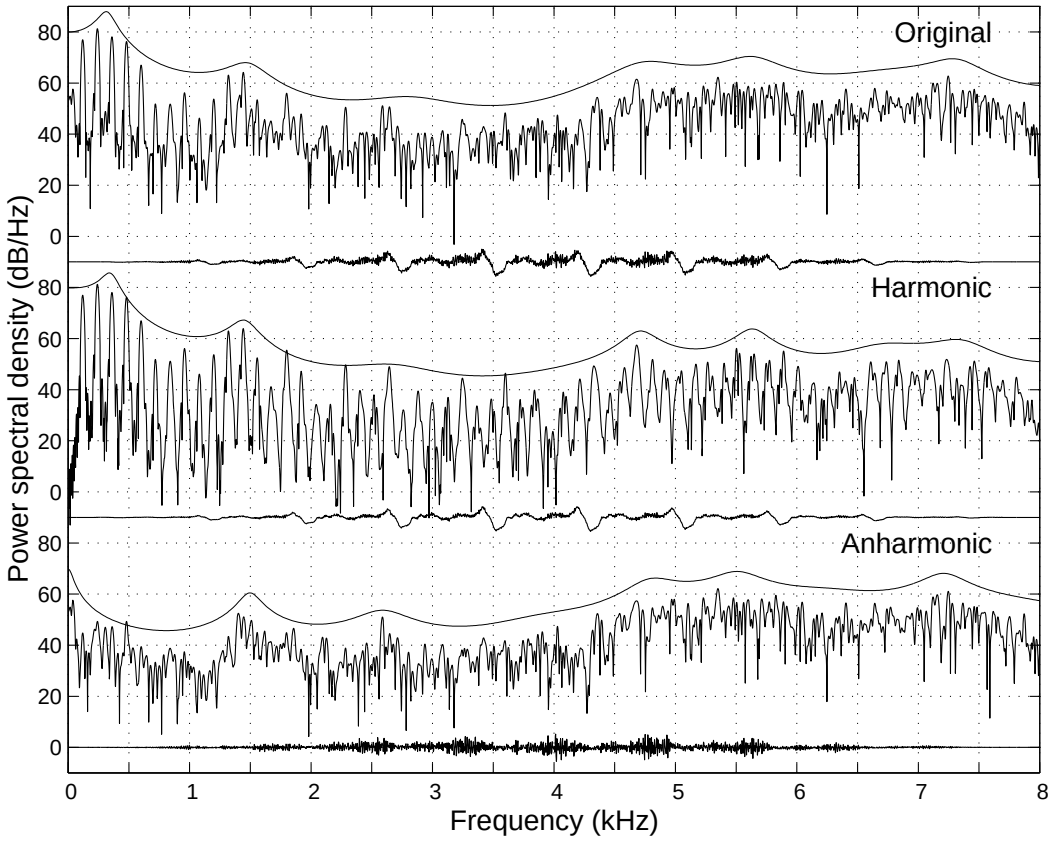


Figure 7.2: Power spectral density (85 ms Hann window centred at 900 ms, $\times 4$ zero-padded, re. 2×10^{-5} Pa) computed from the original signal $s(n)$ (top) for the sustained fricative [z:] by an adult male subject PJ, $^{\#}2$, from the harmonic estimate $\tilde{v}(n)$ (middle) and from the anharmonic estimate $\tilde{u}(n)$ (bottom), whose time series are inset underneath each graph (anharmonic signal magnified by two). The frequency response of the corresponding LPC filter is overlaid above each graph.

anharmonic component dominates, also as expected.

7.1.3 Short-term power (STP)

The envelope of the signal can be described by its RMS amplitude or mean square value, and is itself a slowly varying time signal. Having both the harmonic and anharmonic signals, we can investigate not only the ratio of their envelopes, the short-term HNR, but also their individual trajectories in terms of their short-time power (STP). The STP is a moving, weighted average of the squared signal, centred on time p, which is defined, for any signal $y(n)$, as:

$$P_y(p) = \frac{\sum_{m=0}^{M-1} x^2(m)\, y^2(p+m-M/2)}{\sum_{m=0}^{M-1} x^2(m)}, \tag{7.1}$$

using a smoothing window $x(m)$ of length M. Thus, P_v is the STP of the harmonic component and P_u that of the anharmonic component. The window x acts as a low-pass filter on the squared signals, whose roll-off frequency is governed by the window length M, which reduces the interference from higher harmonics. As such, periodic variations in STP are eliminated with the larger window, yet remain, albeit at a reduced amplitude (-6 dB), with the shorter window.[1] For each computation of the STP, we set M to a constant and used a Hann window: $x(m) = \frac{1}{2}\left(1 - \cos\frac{2\pi m}{M}\right)$ for $m \in \{0, 1, \ldots, M-1\}$, which implies a denominator of 3/8 in Eq. 7.1. In the present study, we were interested in features visible only at high time resolution (of order less than two pitch periods) so, although we were computing the power of the signals, $\hat{v}(n)$ and $\hat{u}(n)$ were used to calculate P_v and P_u, rather than the power-based $\tilde{v}(n)$ and $\tilde{u}(n)$. (Recall that $\tilde{v}$ and $\tilde{u}$ were designed for longer term, narrow-band spectral analysis.) In doing so, we are exploiting the PSHF's signal reconstruction in order to generate features by subsequent (asynchronous) analysis.

The use of these derived measures is best demonstrated with the transition between a vowel and a mixed-source sound that has a strong anharmonic component. Averaging P_v and P_u over a frame comparable with a pitch period, we can see finer variations such as those of the anharmonic component caused by the modulation of the noise, as noted in the vowel-fricative transition [-ɑz-] of the previous chapter (Fig. 6.4). The [ɑz:] vowel-fricative transition example from this chapter (Fig. 7.1) was used to calculate fast and slow STPs, which are plotted in dB in Figure 7.3. To observe fast variations (of the order of a pitch period), the window length was set to the mean period, $M = \langle T_0 \rangle$; for slower variations over the length of the utterance, the window length was set to four times the mean period, $M = 4\langle T_0 \rangle$. The STPs P_v and P_u of the harmonic and anharmonic components, respectively, were calculated for over [ɑz:], and are plotted in decibels in Figure 7.3.

[1] Note that the STP can also be computed in a pitch-scaled way, but there is little advantage from this minor adjustment to the roll-off frequency, for the range of f_0 values within each token.

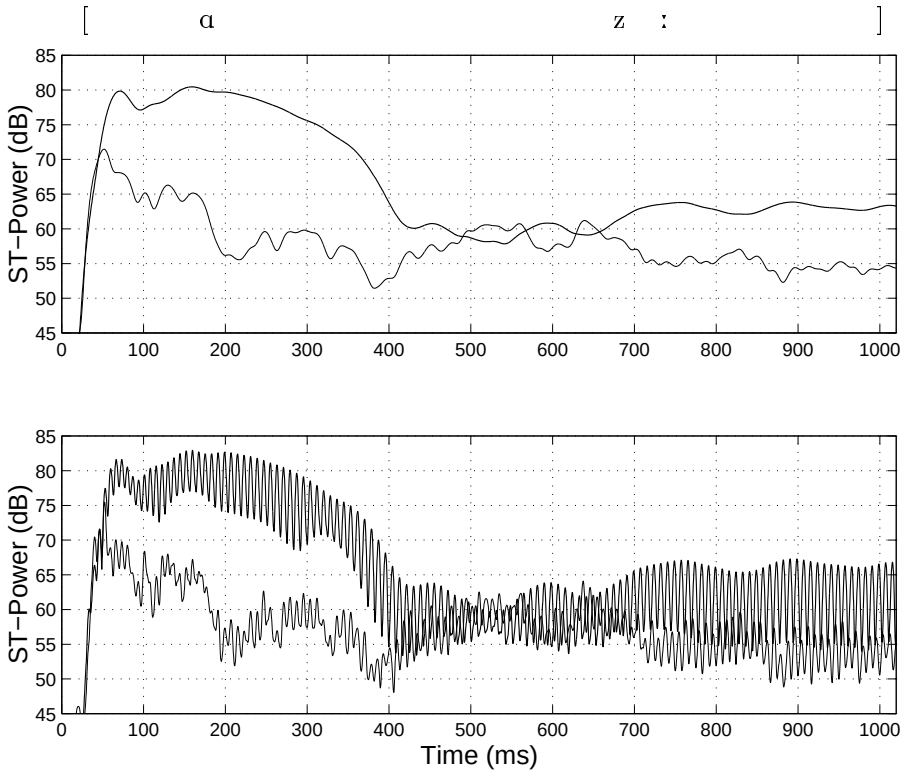


Figure 7.3: The short-term power (STP) for the decomposed components from #2 by subject PJ, calculated for slow ($M \approx 32$ ms, top) and fast ($M \approx 8$ ms, bottom) variations: (thick) harmonic P_v, and (thin) anharmonic P_u. (See time series in Fig. 7.1.)

Figure 7.3 shows the course of the harmonic and anharmonic STPs in decibels, which are smoothed over time. The difference between the harmonic and anharmonic slow STP trajectories ($M = 4\langle T_0 \rangle$, top) is the short-term HNR which, besides voice onset, shows a noticeable change at about 400 ms in the transition from vowel to fricative. Indeed, after voicing has peaked towards the beginning of the vowel (at about 160 ms), the harmonic amplitude dies away, reaching a maximum decay at the transition (circa 400 ms). After some overshoot and subsequent fluctuations it returns to a steady value (c. 700 ms).[2] The anharmonic component grows during the development of the fricative (380–500 ms), undergoes a period of oscillation (500–660 ms) and finally settles down to a reasonably steady value. Note that the fluctuations of the two components at the start of the fricative are roughly equal and opposite. The initial period fluctuations at voice onset cause errors in the harmonic estimate, which get replicated, in negative, in the anharmonic estimate. Otherwise, the HNR is at least +10 dB in the vowel, rising to more than +20 dB at the steadiest point (around 200 ms). In the fricative, values range from −3 dB to +10 dB, settling to about +8 dB in the fully-established part. Their trajectories agree with our earlier observations, but there is evidence of overshoot in the fricative (630–800 ms) before the final equilibrium was reached at c. 860 ms.

The fast STP curves ($M = \langle T_0 \rangle$, Fig. 7.3, bottom), which were computed using the single-period smoothing window, exhibit the same general trends, but have an oscillating element superimposed, which is caused by the modulations in signal power within individual pitch periods. The window x acts on the squared signals effectively as a low-pass filter, whose roll-off frequency is governed by the window length M. As such, periodic variations in STP are eliminated with the larger window, yet remain, albeit at a reduced amplitude (−6 dB), with the shorter window.

7.2 Modulation analysis

In the previous section, we demonstrated the potential for using the PSHF to enable separate analyses of voiced and unvoiced components in mixed-source speech. In this one, we go one step further by relating these parallel analyses to one another.

7.2.1 Pitch-scaled demodulation

To quantify the oscillations in STP, we calculated their magnitude and phase by complex demodulation of the logarithmic signals $10 \log_{10} P_v$ and $10 \log_{10} P_u$ (defined in Eq. 7.1, using

[2] Considering that a more abducted or open glottis would allow a greater air flow, and probably weaken the glottal closure, it is not surprising that the fluctuations of the two components at the start of the fricative are roughly equal and opposite.

$M = \langle T_0 \rangle$). We took pitch-scaled frames of the signal, as for the PSHF ($N = 4T_0$, Hann window w), and extracted the first harmonic, f_0:

$$\dot{P}_y(p) = \frac{10 \sum_{n=0}^{N-1} w(n) \exp\left(\frac{-j8\pi n}{N}\right) \log_{10} P_y\left(p + n - \frac{N}{2}\right)}{\sum_{n=0}^{N-1} w(n)} \tag{7.2}$$

which provided the outputs $\dot{P}_v(p)$ and $\dot{P}_u(p)$ as complex Fourier coefficients, that is the magnitude and phase of the modulation, rather than as reconstructed single-harmonic signals.[3] Note that the tittle (i.e., the dot over the P) thus denotes "the modulation of". Implicit in the demodulation analysis is the assumption that the turbulence-noise source is multiplied by some signal that is related to the vibration of the vocal folds. Thus, by rejecting the higher harmonics, we can take this model as a first order approximation, and extract reliably the phase of the principal mode, that at the fundamental frequency.

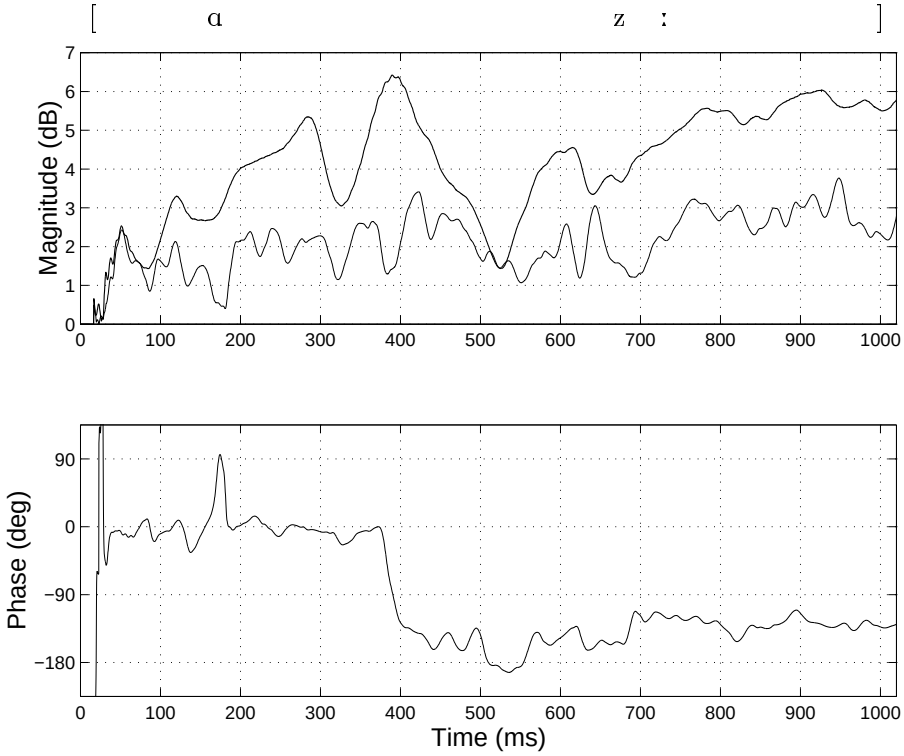


Figure 7.4: Modulation of the STPs ($M = \langle T_0 \rangle$) at f_0 using token #2 by subject PJ (see Fig. 7.1), plotted as magnitudes (top: harmonic, thick; anharmonic, thin) and the phase difference (bottom).

The modulation amplitudes are shown in Figure 7.4 (top) together with the relative phase (bottom). The modulation phases, which continually rotate at approximately the fundamental frequency f_0, are unwrapped and then subtracted from each other to form the phase difference between the modulation of the harmonic component and the modulation of the anharmonic

[3] The advantage of smoothing the STP, before extracting the modulation at f_0, is that it is then less susceptible to interference from higher harmonics.

component, as plotted (bottom). The degree of modulation of the harmonic part (Fig. 7.4 top, thick line) varies considerably during the vowel and the transition, but is more consistent during steady frication. The modulation amplitude is proportionately similar in the vowel and the fricative, and reaches its maximum value right at the transition into the fricative ($\sim$400 ms). It has minima at the points of weak voicing (around 520 ms and 640 ms), but otherwise grows in the fricative towards a steady value of approximately 6 dB. In contrast, the modulation of the anharmonic component is relatively constant throughout, although it is slightly higher at about 3 dB in the steady fricative. There are no clear trends in the vowel; in the fricative, it is arguable whether or not the dips following the points of weak voicing (550 ms and 690 ms) are significant, although quieter phonation might be expected to cause a reduction in the subsequent modulation.

The phase difference (Figure 7.4, bottom), however, gives a more clear-cut picture. During the vowel, the phase difference between the two sets of modulation coefficients is approximately zero, but it changes abruptly at the transition towards a markedly different equilibrium c. $-130°$. We can calulate the mean phase more precisely by considering a series of unit vectors, each with its argument set equal to the instantaneous phase difference, θ:

$$\theta(n) = \arg\left(\frac{\dot{P}_u(n)}{\left|\dot{P}_u(n)\right|}\frac{\dot{P}_v^*(n)}{\left|\dot{P}_v(n)\right|}\right), \tag{7.3}$$

where $\dot{P}_v^*$ is the complex conjugate of $\dot{P}_v$, and $\dot{P}_y/\left|\dot{P}_y\right| = \exp\left(j\arg\left(\dot{P}_y\right)\right)$ is the unit vector with the same phase as the modulation coefficient $\dot{P}_y$, for any y. To avoid phase wrapping errors, unit vectors were used to average the phase in a mathematically-consistent circular algebra. Thus, the (unweighted) time-averaged phase, with its standard deviation, is:

$$\langle\theta\rangle = \arg\left(\vec{e}_\theta\right) \pm \sqrt{\frac{\sum_{n=1}^{S}\left|\exp\left(j\theta(n)\right) - \vec{e}_\theta\right|^2}{S-1}}, \tag{7.4}$$

in radians, where S is the number of sample points, and the mean unit vector $\vec{e}_\theta$ is:

$$\vec{e}_\theta = \frac{\sum_{n=1}^{S}\exp\left(j\theta(n)\right)}{S}. \tag{7.5}$$

For token #2 in Figure 7.4 (bottom), $\langle\theta\rangle = -2° \pm 20°$ during the vowel (40–370 ms), and $\langle\theta\rangle = -128° \pm 8°$ during the fricative (700–1000 ms). This marked difference suggests that more than one voiceless source is in action. The finding is not unexpected yet, as a positive result, it can be used to explore variations in the source interaction quantitatively.

7.2.2 Using EGG as a reference signal

With a view to telling which component is causing the change in the phase difference, we sought to relate the phases to some independent measurement of the glottis. An ideal reference signal would be the glottal waveform itself, but for practical purposes, the glottal area or its electrical

impedance, which can be obtained using an EGG, may be used. Using the coefficient of the EGG signal at f_0, $\dot{L}_x(n)$, we compute the phases of the components:

$$\phi_v(n) = \arg\left(\frac{\dot{P}_v(n)}{\left|\dot{P}_v(n)\right|} \frac{\dot{L}_x{}^*(n)}{\left|\dot{L}_x(n)\right|} \right) , \tag{7.6}$$

$$\phi_u(n) = \arg\left(\frac{\dot{P}_u(n)}{\left|\dot{P}_u(n)\right|} \frac{\dot{L}_x{}^*(n)}{\left|\dot{L}_x(n)\right|} \right) . \tag{7.7}$$

Ignoring the effect of phase wrapping, the phases can be subtracted to give Eq. 7.3: $\theta = \phi_u - \phi_v$.

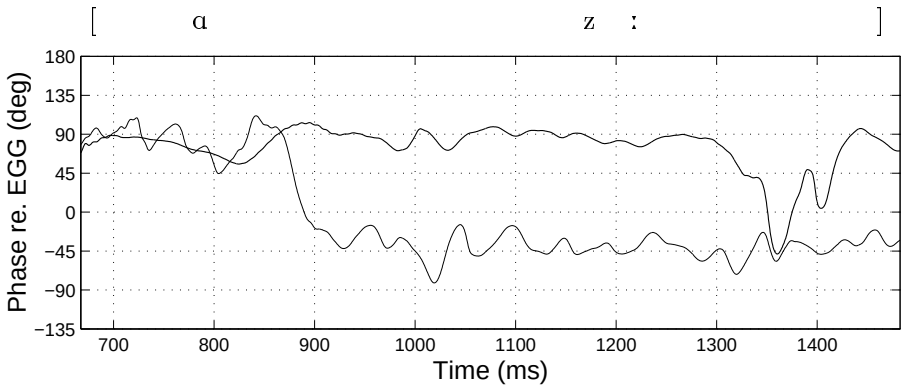


Figure 7.5: Phase of the harmonic (thick) and anharmonic (thin) modulation components for *C5*-[ɑz:] (#3) by subject PJ, related to that of the simultaneously-recorded EGG signal.

Figure 7.5 contains the phase trajectories of the two components for another [ɑz:] token, #3, spoken by subject PJ, which do not exhibit the overshoot phenomenon that we saw earlier (Fig. 7.3 top). Both phases hover close to +90° initially. The harmonic component is perturbed near the transition, returning to approximately the same value for the fricative, except when it strays as voicing momentarily falters (between 1300 ms and 1430 ms).

The anharmonic component shows greater variability, but approaches an equilibrium value after the transition that is distinctly offset from the average during the vowel. The change noted in $\langle\theta\rangle$ thus appears to be due primarily to changes in ϕ_u, signalling a change in source mechanism for the unvoiced component. We expect that the anharmonic component during the vowel is due to a slight breathiness, i.e., turbulence noise generated in the vicinity of the glottis, and that during the following [z:], the anharmonic component is primarily due to turbulence noise generated downstream of the tongue-tip constriction. The step change in ϕ_u at the vowel-fricative transition therefore corresponds to a change in source location. This effect would predict that the amount of phase change should depend on the fricative's place, which we will investigate in Section 7.3. It should be noted that a phase difference of approximately zero could as easily be the product of perturbation errors (e.g., from jitter and shimmer) in the processing as of an in-phase modulated noise source. Nevertheless, examination of the time-series signals for the harmonic and anharmonic components for over twenty examples gives

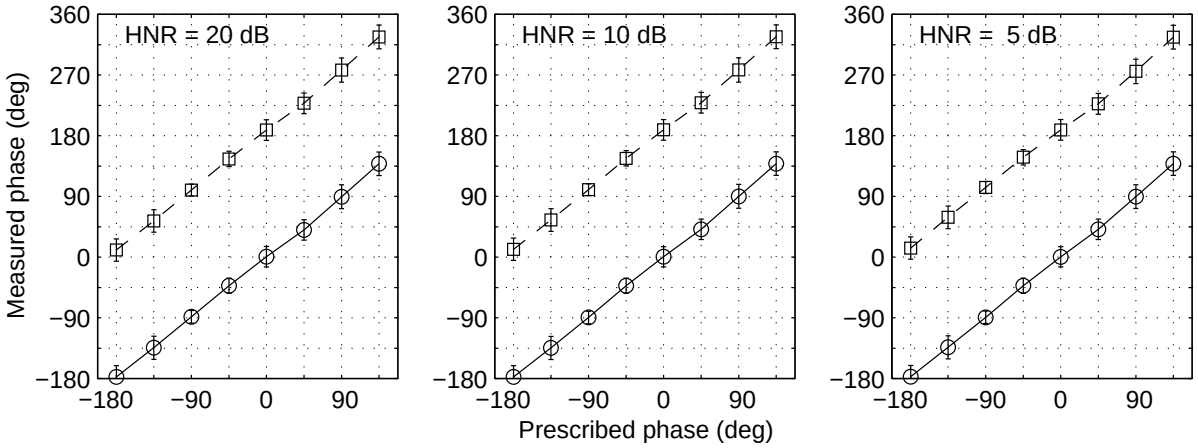


Figure 7.6: Phase measured by the demodulation analysis against that prescribed. Eight tests were conducted at 45° intervals for three HNR levels: (left) 20 dB, (centre) 10 dB and (right) 5 dB. The circles joined by a solid line were referred to the glottal pulses, while the squares on the dashed line were to the harmonic component.

us confidence that the STP, as a summary of signal amplitude (or envelope), contains useful information about the sources.

7.2.3 Validation of phase estimate

Now that we have introduced a quantitative technique for measuring the PSHF's property of maintaining the anharmonic component's envelope, let us perform an evaluation using the synthetic signals from Chapter 5. This time we set the angle β in Eq. 5.23 to a range of values and assess the ability to estimate that angle from the phase of the modulation of the anharmonic component. Therefore, using the procedure outlined above we estimated the phase offset β for each of its eight specified values (0°, 45°, 90°, etc.) at three HNRs (20, 10 and 5 dB). The results are plotted in Figure 7.6. All modulation phases measured from the decomposed synthetic signals were within 5° of their specified values. The mean error was less than 1° and the inter-measurement standard deviation was 2°. There were no noticeable differences across the different HNR levels, except perhaps a slight trend in the (much higher) intra-measurement deviations, which were 15°, 13° and 13°, respectively.

7.3 Results

Following decomposition of a variety of voiced fricatives, modulation analysis of the components was performed, with phase referred to the EGG signal. Results were attained for a wide range of constriction locations with approximately constant f_0, and for a few of these with changing f_0.

7.3.1 Sustained fricatives

The magnitude and phase of the modulation coefficients were determined for 10 fricative tokens that included seven different places of articulation. All of the tokens were similarly pitched at $f_0 = 120 \pm 5\,\mathrm{Hz}$, and sustained by subject PJ for at least 4 s, of which a steady section of approximately 1 s duration was analysed. For some cases, the section analysed included a part of the contextualising vowel; for others, only the fricative was included. The PSHF was used to decompose each example, and modulation coefficients of the harmonic and anharmonic components were calculated, as described in the previous section. Finally, the coefficients were averaged over the fricative, excluding periods of devoicing, vowel-fricative transitions and two pitch periods from either end of the section. The time-averaged magnitudes and phases are plotted in Figure 7.7. The points plotted on the vertical grid lines were all from steady regions of voicing, whereas those adjacent suffered an interruption in voicing. The latter are discussed in Section 7.4.6, below.

As mentioned in Section 7.2.1, the magnitudes (Fig. 7.7, top) were all halved by the low-pass effect on signal power of the windowing, which was adjusted accordingly for each measurement to allow comparisons between harmonic and anharmonic STP, and across different phonemes. The magnitude of the modulation of the harmonic components (thick) is $3 \pm 1\,\mathrm{dB}$ and, in all but one case, is greater than that of the anharmonic components (thin). The anharmonic modulation magnitudes were equally variable, but ranged from almost zero in the bilabial fricative [β] to 2 dB in [z] (the same as that of the harmonic modulation).

The phase of the modulation coefficients was referred to the EGG signal by subtracting the phase of its f_0 component, as before. Care had to be taken in aligning pitch, power (STP) and phase vectors in the analysis, but the difference between using the pitch extracted from the acoustic signal versus that from the EGG was found to be negligible. The unweighted-mean values are plotted in Figure 7.7 (bottom) with error bars indicating one standard deviation (± 1 s.d.), time-averaged over the appropriate portion of the token. Of the two components, the harmonic results showed greater consistency within each phase measurement; across measurements, these values were all in the vicinity of $+100^\circ \pm 20^\circ$. The anharmonic phases, although more variable, were all distinct from their harmonic phases, except for [ʕ]. Moreover, where the transition from the vowel was included in the analysis segment, a clear step was seen in the time series of the anharmonic modulation phase.

The phase of the modulation of [β]'s anharmonic component had the largest variance, which was related to the unusually small amount of modulation and rendered it most susceptible to interference from disturbances. Since the anharmonic modulation in [β] was therefore poorly correlated with the EGG, we shall ignore this phoneme in subsequent evaluation. For the remaining anharmonic phase data, there were two notable trends: (i) the mean phase increased

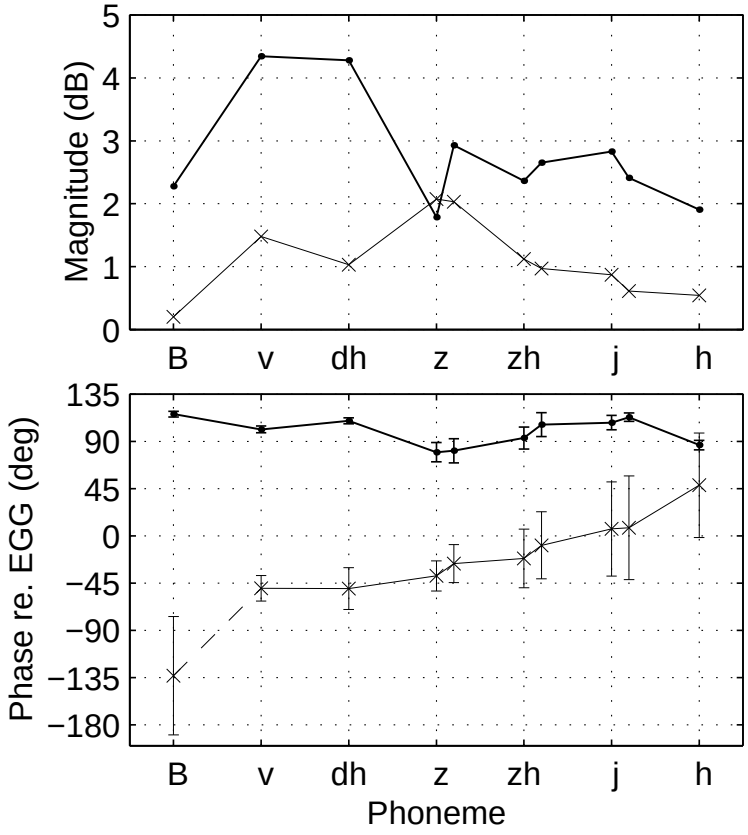


Figure 7.7: Magnitude (top) and phase (bottom) of modulation coefficients, referred to the EGG signal, versus place of articulation for sustained fricatives [β, v, ð, z, ʒ, ɣ, ʕ] by subject PJ. Harmonic (•, thick line) and anharmonic (×, thin line) components were plotted with (±1 s.d.) error bars. Those measurements on vertical grid lines are for normal voicing; those adjacent (to the right), where a pair of measurements are shown, were taken from a section that had been interrupted by devoicing.

as the place of constriction moved in a posterior direction, and (ii) so did the variance. The systematic change of phase with place seems worth further investigation, although we might well expect the phase to depend also on f_0. Any delay in the system, such as the propagation time from the lips to the microphone, would add a phase term that increased linearly with f_0, its gradient dependent on the amount of delay. In the following section, we investigate the relationship between the pitch and anharmonic phase during sustained fricatives that contain changes in f_0, and attempt to identify the cause of any delays.

7.3.2 Pitch glides

When using spot measurements of phase for determining delay times, the main concern is that phase wrapping may occur, e.g., a phase reading of 420° might be misinterpreted as only 60°, or vice-versa. The number of cycles is important because long delays, i.e., greater than a period, inherently entail phase wrapping. A simple test for phase wrapping can be carried out by altering the fundamental frequency f_0 and by noting the phase changes. A few spot measurements can be made or, more dependably, a continuous measurement during a pitch glide. For a constant delay τ_u, the phase is simply a linear function of frequency:

$$\phi_u = 2\pi\tau_u f_0 + \beta\,, \tag{7.8}$$

where β is the phase offset between the actual modulating signal, whatever it may be, and the EGG signal. The phases ϕ_u and β can take any real value, although in our initial measurements they lie in the range $\pm 180°$. Hence, provided other independent variables remain unaltered, the gradient of the phase with respect to frequency provides an absolute estimate of τ_u, the delay duration for a given phoneme.

Subject PJ was asked to sustain a fricative during a smooth pitch glide sandwiched between two notes about a perfect fifth apart. That is, a constant-f_0 fricative was held for at least 1 s, then f_0 was increased steadily to approximately $1.5f_0$ over a similar period, and finally the fricative was held at the higher note of about $1.5f_0$ for at least another second, taking about 5 s in total. Recordings were also made of descending pitch glides.

For all of the tokens analysed, the time series of the anharmonic modulation phase showed a definite correlation with the extracted f_0, and both parameters exhibited distinct equilibria at the end conditions, which were connected by a gradual transition. The relationship between f_0 and the phase ϕ_u can be seen more clearly by plotting them against each other, independently of time. Thus, Figure 7.8 is a scatter diagram of the anharmonic STP modulation phase versus fundamental frequency for the sustained fricative [z:], during a descending pitch glide.[4]

So as to estimate the values of the constants in Eq. 7.8, τ_u and β, from these data, we have fitted a regression line that represents the response of a noise modulated with a constant delay.

[4]Every one in ten points has been plotted, so the values have been effectively sampled at 4.8 kHz.

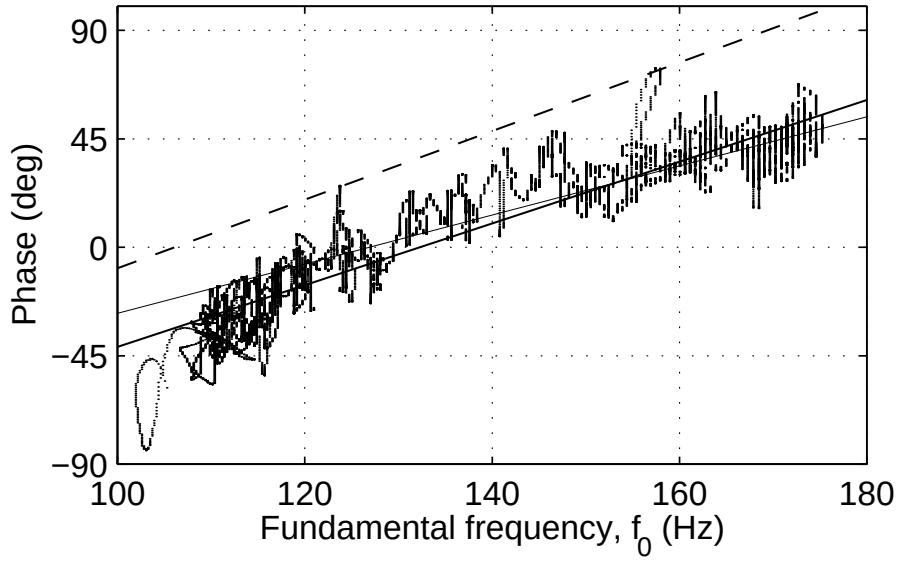


Figure 7.8: Scatter plot of the anharmonic modulation phase versus fundamental frequency for the sustained fricative [z:] by subject PJ during a descending pitch glide, with its regression (thick solid line), and those of an ascending [z:] (thin solid line) and a descending [ʒ:] (thick dashed line).

However, f_0 is not the only parameter adjusted during the glide, and changes in flow rate and open quotient, for instance, may account for some of the variations beyond measurement error. These uncertainties would also suggest that we avoid using a higher-order model for risk of over-fitting. Nevertheless, in this example, the points do lie roughly along a straight diagonal line, in the range $\pm 45°$, except for a few stray excursions that occurred at transitions or near a singularity, where the modulation amplitude was almost zero. There is a higher density of points at either end of the trajectory line due to the period of constant pitch before and after the frequency ramp. The deviation from this line, $\sigma \approx 10°$, is of the same order as the deviation of the (constant-f_0) sustained fricatives considered earlier. Owing to the integer quantisation of the extracted pitch period (in sample points), the fundamental frequency values also exhibit quantisation, which explains why the data points lie on a set of vertical lines.

The best-fit line (thick solid line in Fig. 7.8) was calculated for the plotted data points by a least-mean-squares regression and provides good general agreement. The gradient provides an estimated delay time of $\tau_u \approx 3.8$ ms, and the intercept with the y-axis at $f_0 = 0$ was $\beta \approx -170°$. Regression lines were also calculated for two other examples: [z:] ascending and [ʒ:] descending. The lines for [z:] are within 10° of each other for the ranges of f_0 measured, although their gradients differ, which suggests that some other factor may have influenced these results. The line for a descending [ʒ:] is set apart from those for [z:], but has a similar gradient, particularly to that of the descending [z:].

Phoneme	f_0 (Hz)	τ_u (ms)	β (°)	σ (°)
[z:] ascending	$125 \rightarrow 175$	2.8	-129	10
[z:] descending	$111 \leftarrow 172$	3.8	-169	11
[ʒ:] descending	$121 \leftarrow 178$	4.0	-154	22

Table 7.1: The anharmonic delay τ_u, the offset phase β and the standard deviation σ about the corresponding regression line, for three f_0 glides by subject PJ.

The values of β and τ_u for all three cases are listed in Table 7.1, with the mean values of the f_0-glide endpoints. The difference between the two descending fricatives [z:] and [ʒ:] was as expected in both direction and scale, yet there was a considerable discrepancy between the values calculated for the ascending and descending [z:], which was exacerbated by the extrapolation to $f_0 = 0$. Given that the propagation time for an acoustic wave from the lips to the microphone is 2.9 ms ($r = 1\,\mathrm{m}$, $c_0 = 343\,\mathrm{m/s}$, room temperature, dry air) and acoustic propagation in the tract would take about 0.5 ms ($l = 16.5\,\mathrm{cm}$, $c_0 = 359\,\mathrm{m/s}$, body temperature, saturated air), the times derived from the gradient are of an appropriate order of magnitude. The zero-frequency phase offset β, despite these errors, corresponds to a point between one-half and three-quarters of the way through the open portion of the glottal cycle. We shall speculate about potential interpretations of the coincidence of this timing relationship with the maximum glottal flow in the following section. For fricatives showing a higher variance, the scatter plots are less informative. Critically, no phase wrapping of the modal trajectories took place for any of the fricatives examined, which validates the order of our earlier phase measurements.

7.4 Discussion

We would like to convert the reported phase values into delay times in order to relate a peak in the acoustic response to the event that caused it. Later in this section, we attempt to explain the pattern of delays for the different fricatives in terms of a possible aero-acoustic mechanism of sound production.

7.4.1 From phase to delay

The glottal closure is commonly assumed to give the principal acoustic excitation of the vocal tract. The harmonic component $v(n)$ should then consist primarily of the vocal tract response to that excitation. The smoothed STP of $v(n)$ has a peak every cycle that is slightly delayed with respect to the instant of excitation, and further delayed owing to the acoustic propagation time from the glottis to the microphone in the far field. We computed its phase ϕ_v with respect to the peak of the fundamental component of the EGG signal. To refer it instead to the moment

of closure of the vocal folds, we subtract $\alpha = \arg\left(\dot{L}_x\right)_{\rm cl}$; to convert this phase to a time delay, we divide by the instantaneous fundamental frequency:

$$\tau_v = \frac{\phi_v - \alpha}{2\pi f_0}\,, \tag{7.9}$$

where ϕ_v is defined by Eq. 7.6. The anharmonic component $u(n)$ consists primarily of the vocal tract response to the noise excitation. We wish to convert ϕ_u to a time delay also, but it is not clear whether we should refer ϕ_u to the same instant of closure of the EGG signal. If we use the same angle α as in Eq. 7.9, we are effectively assuming a model of the modulation mechanism, namely that the peak amplitude of the turbulence noise source is evoked by the excitation originating from the instant of glottal closure. We wish instead to deduce the mechanism controlling the modulation, by using the phase difference expressed as a time delay. Therefore, to refer the phase to an unknown point in the EGG signal, we subtract the angle β:

$$\tau_u = \frac{\phi_u - \beta}{2\pi f_0}\,. \tag{7.10}$$

where ϕ_u is defined by Eq. 7.7. For our initial discussions, we set $\beta = \alpha$.

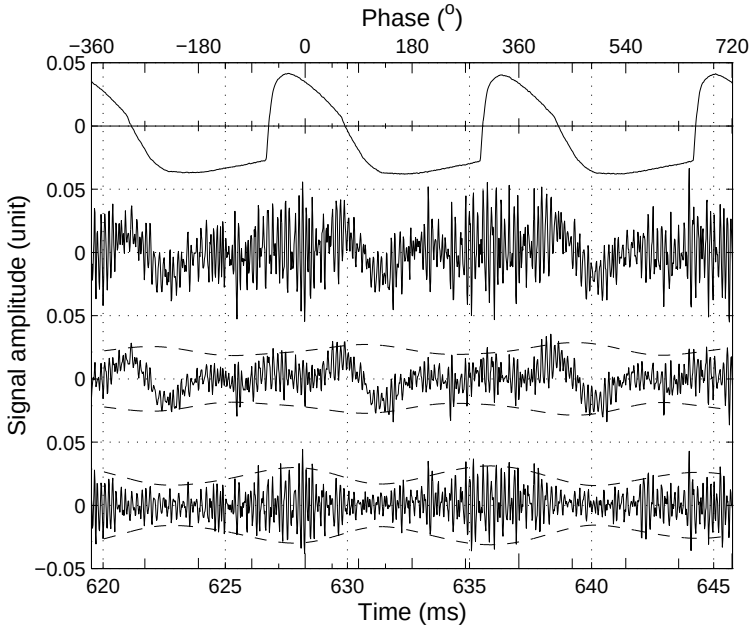


Figure 7.9: Time series during a sustained [z:] by subject PJ: (from top) EGG signal L_x, sound pressure s, harmonic part v, and anharmonic part u.

Figure 7.9 shows a set of four synchronous time-series signals during the fricative [z:] sustained by subject PJ, which are (from top) recorded EGG, $L_x(n)$, recorded sound pressure, $s(n)$, and the decomposition into the harmonic and anharmonic signals, $\hat{v}(n)$ and $\hat{u}(n)$. The dashed lines around the harmonic and anharmonic components represent their envelopes (i.e., $\pm 2\sqrt{P_v}$ and $\pm 2\sqrt{P_u}$). The EGG measures the time-varying (high-pass filtered) part of the trans-glottal conductance, which is at a maximum when the glottis is closed. It shows a sharp

rise at the instant of closure, occurring at around -0.4π ($-72°$), with respect to the EGG signal's fundamental component, whose phase is indicated by the upper abscissa in Fig. 7.9. This phase offset is slightly less than a quarter of a cycle, because of the long open portion and the abruptness of the closure. Although the phase may change slightly throughout the recorded corpus and for subjects other than PJ, the value of $\alpha = -0.4\pi$ shown here is used in all cases to refer the harmonic component to the same instant of the EGG signal.

Phoneme	v	ð	z	ʒ	ɣ	ʕ	ɑ$^{\text{h}}$
Distance (cm)	0.0	0.4	1.1	2.2	5.2	10.3	12.9

Table 7.2: Estimated distance from the constriction to the teeth for sustained voiced fricatives by subject PJ.

Through a separate study (Shadle et al. 1999), we obtained magnetic resonance imaging (MRI) data for subject PJ, saying [p$^{\text{h}}$ɑsi]. Combining these with articulatory phonetics, we were able to estimate the constriction location for each phoneme. Distances along the vocal tract were measured from the glottis, and the position of the teeth was estimated in relation to the lips and the hard palate (upper) or tongue body (lower). Table 7.2 lists all the constriction-teeth distances, which agree closely with Table I in Narayanan et al. (1995). For the breathy vowel [ɑ$^{\text{h}}$], the place of greatest constriction was assumed to be the glottis.

Ideally, we would like to characterise each phoneme by two distances: from glottis to place of constriction, and from constriction place to the location of turbulence noise generation. Different aspects of sound generation take place over these two 'paths'. While for some fricatives it is well known that noise generation is highly localised at the teeth (e.g., [s, ʃ, z, ʒ]), for others the noise source appears to be distributed, for instance, along the hard palate for [ç] (Shadle 1991). The distance from the constriction to the source location is thus less precisely known for some fricatives. All delays are therefore calculated using the constriction-teeth distances given in Table 7.2. These values were used for all three subjects, regardless of minor inter-subject variation in physical dimensions. Although women's vocal tracts are generally shorter than those of men, most of the difference is in the pharynx. Since, for LJ and SB, we are dealing with distances from within the oral cavity to the teeth, the variation is considered negligible. Although this part of the procedure is crude compared with the signal processing, it enables us to visualise our results in a way that has greater physical meaning. Bearing in mind that the teeth will not necessarily be the source location in all cases, we can nevertheless interpret trends and make order of magnitude calculations to help indicate the aero-acoustic processes that are likely to be operating.

The delays calculated for the voiced fricatives of three subjects are plotted against place of articulation in Figure 7.10, including one breathy [ɑ]-vowel (PJ). For reference, the lip-

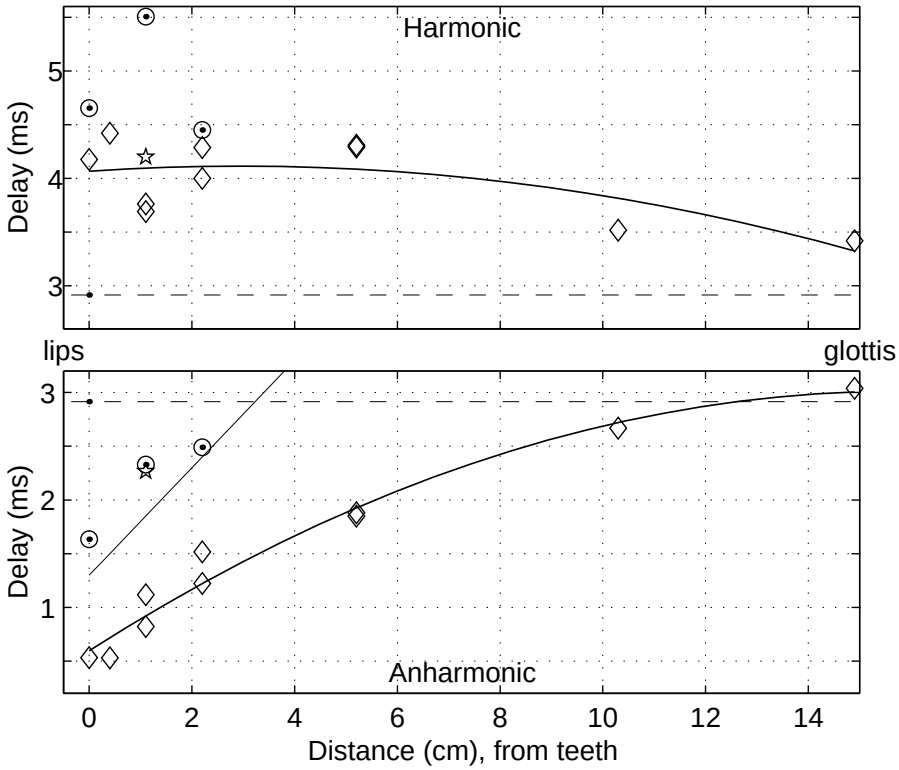


Figure 7.10: Harmonic and anharmonic delay times, τ_v (top, Eq. 7.9) and τ_u (bottom, Eq. 7.10) respectively, versus distance of constriction from teeth, for subjects PJ ($\diamondsuit$), LJ ($\odot$) and SB ($\star$). The dashed line is the predicted lip-microphone propagation delay τ_R, the thin solid line is the predicted total delay, and the thick solid line is the quadratic line of best fit.

microphone propagation time is shown as a dashed horizontal line, $\tau_R = 2.9$ ms for a microphone at 1 m (speed of sound $c_0 = 343$ m/s).[5] In Figure 7.10 (top), the delay times τ_v are all greater than the acoustic propagation delay, as expected. The additional delay, the reverberation lag, is reasonably consistent across phonemes, showing a mean value of 1.3 ms and no significant trend. In contrast, τ_u (Fig. 7.10, bottom) is generally below τ_R. Since the largest portion of these delays is, in fact, the wave propagation time from the lips to the microphone (which is obviously identical for both components), any variations in the delay are attributable to other causes. Such causes include jitter/shimmer effects, changes in glottal waveform, changes in vocal-tract configuration, the measurement noise on the data, processing errors, and actual changes in the source characteristics.

In Figure 7.10 (bottom), the dominant trend for subject PJ is for the anharmonic delay τ_u to increase with distance from the teeth by 0.3–0.5 ms/cm. The anterior results for subject LJ exhibit the same trend, fitting the predicted delay line very closely, and the single value obtained for our female subject SB lies inbetween those of the male subjects, PJ and LJ. However, before we attempt to interpret the anharmonic τ_u readings in Figure 7.10 (bottom), let us consider the physical mechanisms that could lead to modulation of the frication source, as has been observed.

7.4.2 Theory

For the voiced component, $v(n)$, the instant of glottal closure is commonly assumed to give the principal excitation of the vocal tract. The classic source-filter model predicts that the glottal waveform is introduced to the vocal tract through the larynx, reverberates to and fro within the vocal tract, which adds ripple from the formant resonances, and radiates from the lips, where the velocity waveform is effectively differentiated, as a result of the radiation impedance. The voiced component is therefore dominated by the ringing of the vocal-tract resonances after the instant of glottal closure, which is when the derivative of the glottal flow is at its maximum amplitude over one pitch pulse. Accordingly, the peak in the smoothed STP occurs shortly after the effects of the glottal closure have first reached the observer, because of the resonances. To verify this, we should calculate the harmonic delay time τ_v from glottal closure to peak STP and expect values equal to the propagation time plus a small reverberation lag, as above.

In contrast, the unvoiced part, $u(n)$, is produced in the presence of net flow by a jet that generates a turbulence-noise source, which is somehow modulated by the oscillation of the vocal folds. The sound that is produced reverberates up and down the vocal tract from the source location, whether it be at the constriction, in the jet wake, or at an obstacle downstream, and finally radiates from the lips, as before. We know, from the result in Section 7.3, that phonation

[5] Suitable substitutions for ϕ_v and ϕ_u were made into Eqs. 7.9 and 7.10 to derive the error bars.

by some mechanism induces pulsation of the turbulence noise generated near the supraglottal constriction. As mentioned previously, the complex demodulation we have performed is effectively based on the assumption that whatever the underlying physical mechanism, modulation is occurring in a strict mathematical sense. That is, the envelope of the noise is multiplied by the modulating signal, a sinusoid synchronised to voicing via the EGG signal; the turbulence noise acts as the carrier of this signal. Thus, the pulsed flow velocity is not differentiated by the effect of radiation from the lips, as such, but the (modulated) turbulence noise is; not the modulating signal (or the noise envelope), but the carrier signal, that is, the noise itself. Therefore, to observe the timing characteristics of the modulation, which will help to explain how the turbulence noise is created, we should refer the envelope of the unvoiced component to the glottal flow, for which we can use the STP. Some potential mechanisms will be discussed later in Section 7.4.4. Ideally, we would calculate the time from the peak in glottal flow to the peak in the magnitude of the anharmonic STP. Unfortunately, it is not possible to measure the glottal flow in vivo, so we used a simple model to predict the instant of peak flow using the EGG signal. An idealised glottal flow U_G was generated by fitting a cubic function (such as in Klatt 1987) to the open portion of the cycle, as defined by the peak negative-going change in gradient and the positive-going zero-crossing in the EGG signal.

The time at peak flow also depends on the proportion of the cycle for which the glottis is open, the open quotient, which was typically OQ ≈ 0.65 in our recordings (as in Fig. 7.9). Recalling our assumption that the peak flow was two-thirds of the way through the open portion, the phase of the EGG signal was $\beta = -(2\pi \times 0.65 \times 1/3) - 0.4\pi \approx -0.83\pi$ (-150°). Note that, despite assuming the degree of skewness a priori, the predicted value of β lies well within the range of values estimated from the pitch glide measurements.

7.4.3 Travel times

The production of voiced fricatives involves vibrating vocal folds and a constriction in the supraglottal tract that produces a jet. We would like to know precisely how phonation causes the pulsing of the frication source. It is known from studies using physical, flow-duct models in sound production experiments (e.g., Shadle 1985), that the presence of an obstacle in the path of turbulent flow can enormously enhance the radiated sound. In fact, new sound sources are generated at the obstacle, which have a higher acoustic efficiency. Aero-acoustic theory describes the source found in the free jet and at the obstacle as flow-quadrupole and flow-dipole respectively, predicting greater efficiency for the dipole sources. In speech, particularly for the front, or anterior, fricatives (viz. labial, dental, alveolar), it is known that the teeth play an important role in sound production, generating and shaping the noise formed from any jet that impinges upon them. The path that the flow perturbation must take from glottis

to far-field microphone can be divided into three sections: from glottis to constriction exit; from constriction exit to the principal location of turbulence noise generation; thence to the microphone. The first two paths are the most important with regard to the mechanism of noise modulation, and we can assume that the sound radiated from the lips to the microphone travels acoustically.

Convection involves hydrodynamic fluid motion of flow structures along the tract to the source location, and may be generated at the glottis or any supraglottal obstacle or constriction. These structures convect in one of two forms: as pulsatile bulk flow inhomogeneities or as regions of vorticity. Flow inhomogeneities are unstable structures, which tend to disperse over long distances. Rather like breaking waves in the sea, pockets of slower flow are caught up by faster regions until finally the pressure gradient is unsustainable, the wave breaks and energy is dissipated. Rotational flow, on the other hand, can be transmitted by convection through the fluid in a much more stable way, e.g., as vortex rings. Vortices arriving at a constriction downstream would generate sound by re-attachment of separated flow into the laminar regime. In either case, we would expect the front fricatives, such as /v, ð/, to exhibit weaker modulation than those nearer to the glottis, like /ɣ/.

/z/		t_2 (ms)			/ʒ/		t_2 (ms)			/ɣ/		t_2 (ms)		
		ac	co_1	co_2			*ac*	co_1	co_2			*ac*	co_1	co_2
t_1 (ms)		0.06	1.90	0.63	t_1 (ms)		0.08	3.0	1.0	t_1 (ms)		0.17	6.0	2.0
ac	0.38	0.44	**2.3**	**1.0**	*ac*	0.35	0.44	**3.4**	**1.4**	*ac*	0.27	0.44	**6.3**	**2.3**
co_1	690	690	690	690	co_1	640	640	640	640	co_1	490	490	490	490
co_2	230	230	230	230	co_2	210	210	220	210	co_2	160	160	170	160

Table 7.3: Estimated travel times (ms) for /z/ ($l_1 = 14.6\,\text{cm}$, $l_2 = 1.1\,\text{cm}$), /ʒ/ ($l_1 = 13.5\,\text{cm}$, $l_2 = 2.2\,\text{cm}$) and /ɣ/ ($l_1 = 10.2\,\text{cm}$, $l_2 = 5.2\,\text{cm}$), by acoustic propagation *ac* or by convection *co*, using $U_1 = 200\,\text{cm}^3/\text{s}$ and $U_2 = 600\,\text{cm}^3/\text{s}$ for co_1 and co_2 respectively. The column under t_1 gives the travel times over path 1, and the first row under t_2 those for path 2. The nine values inside each sub-table are $t_1 + t_2$, rounded to two significant figures; those in bold face best match the measured data (see text).

During phonation, the pulsing jet of air exiting from the glottis generates sound and sets up vortical motion. The sound wave travels downstream at the speed of sound; the vortices convect at the order of the mean flow velocity, which is much slower than the speed of sound c_0 (Barney et al. 1999). The effects of phonation, therefore, traverse the first section of the path in two different ways, with two different travel times. The longer that section is, i.e., the more anterior the constriction, the bigger the discrepancy in time will be. The travel time for a sound wave over this first glottis-to-constriction path of length l_1 can be estimated as $\tau_1|_{ac} = l_1/c_0$.

Values are shown in Table 7.3 computed for three different l_1 values ($c_0 = 359\,\mathrm{m/s}$). The convective travel time is estimated as $\tau_1|_{co} = l_1/(V/2)$. A minimum and maximum convective velocity are computed using volume velocities of 200 and $600\,\mathrm{cm}^3/\mathrm{s}$, and an average cross-sectional area through the back cavity of $5\,\mathrm{cm}^2$. It is clear from the values shown in the table that even the lower of the convective delay estimates (co_2) is two orders of magnitude higher than the measured delays. Such delays would be easily observable at any transition, and would in particular lead to extensive phase wrapping on the pitch glides. Further, we observe longer delays (longer by approximately 1 ms) for a more posterior place, whereas a convective mechanism for path 1 would mean that delays would shorten by 50 to 150 ms. Therefore we conclude that the aspect of phonation that modulates the noise travels at the speed of sound over path 1.

The second path extends from the constriction to the principal location of turbulence noise generation. The flow velocity increases in the constriction; at the exit, a turbulent jet forms. The self-noise (from mixing) of the jet is relatively weak for vocal-tract dimensions and flow rates but, whatever obstacle the jet encounters (whether the palate or the teeth), additional turbulence noise is generated that is louder (and can be much more localised). If the jet emerging from the constriction is pulsing, the turbulence noise generated by it will likewise fluctuate, but an acoustic field can also influence the formation of turbulence (Crow and Champagne 1971). We could further consider whether an acoustic field could influence not only the jet structure, but the sound generation where it impinges on the obstacle.

For path 2, we can again make order-of-magnitude estimates of the travel time at acoustic and convective velocities. We estimate l_2 to be the constriction-teeth distance, although we expect that the teeth do not act as the obstacle in all these cases. Again, two values of l_2 are chosen that correspond to the two values of l_1, that is, result in the same vocal tract length in both cases. The acoustic delay is then computed as $\tau_2|_{ac} = l_2/c_0$, as shown in the table. For the convective delay, V is recomputed using a typical constriction area of $0.1\,\mathrm{cm}^2$ rather than the $5\,\mathrm{cm}^2$ used earlier. The same minimum and maximum volume velocities are used, giving much higher values of V.

From Figure 7.10 (bottom), lengthening l_2 from 2 to 5 cm actually increases the anharmonic delay by approximately 0.7 ms. This is consistent with the convective delay computed using the maximum convective velocity (column co_2 in Table 7.3). If travel times were at speed of sound in both paths, there would be virtually no difference in the delay with place. Therefore, the second path must involve some mechanism that convects.

7.4.4 Source modulation mechanisms

What theoretical models exist that describe the modulation mechanism itself? Most of the methods in the literature, summarised in Chapter 1, incorporate modulation by a parameter related to glottal flow, such as the instantaneous component of the volume velocity at the constriction exit, but do not allow for a non-acoustic mechanism, i.e., for propagation velocities other than the speed of sound. The differences with place that we observe in the phase of the anharmonic component are not consistent with models depending only on acoustic propagation.

We have not so far discussed the extensive literature examining interaction of the glottal waveform with the vocal-tract driving-point impedance. Rothenberg (1981) showed, theoretically and by inverse-filtering speech, that the first formant frequency F1 affects the degree of skewing of the glottal waveform U_G: the vowel [ɑ], with its high F1, has a more skewed U_G (peak U_G occurring later in the glottal cycle) than does [i], with low F1. Since all of the English voiced fricatives have lower F1 than [ɑ], the peak U_G is predicted to shift earlier in the cycle during [ɑF], which was borne out by Bickley and Stevens' results (1986) for consonantal constrictions at the lips. Nevertheless, though such a mechanism could perhaps explain why the phase difference changes during the vowel-fricative transition, it does not explain the amount of change we observe (ranging from 40° to 150°) nor the difference with place, which should affect F2 and higher formants rather than F1.

Crow and Champagne (1971) showed that acoustic excitation applied to air in a duct upstream of the jet nozzle could induce an orderly structure in the jet wake, with a preference for $St = fD/V = 0.30$. Such a structure appears when the acoustic velocity is greater than 1 % of the mean flow speed V at the nozzle exit (nozzle diameter D). The turbulence noise spectra show that the forcing has the effect of suppressing background noise and enhancing noise at frequencies near the forcing fundamental and its harmonics.

We cannot compare all aspects of Crow's and Champagne's results to ours because the relevant vocal-tract parameters cannot be measured accurately enough. However, we estimate that Strouhal numbers for voiced fricatives range from 0.3 to 0.9, based on $f = f_0$, a typical constriction diameter D, and the volume velocities U used in Table 7.3. The forcing takes some (unspecified) time to alter the shape of the jet; any change in the jet travels downstream at its convection velocity. We conjecture that the sound generation mechanism with which we are chiefly concerned, that of the jet impinging on an obstacle, would, in the presence of the 'forcing function' of phonation at f_0, exhibit non-linear emphasis of f_0 and its harmonics, similar to the free jet spectra shown by Crow and Champagne. Any change in f_0 would affect the noise generated after a delay, related to the convection velocity and the distance from constriction to obstacle. Their results provide a plausible mechanism for the modulation of voiced fricatives, but do not help us to estimate β, the angle that determines the phase of the glottal cycle to

which we should refer the modulation of the anharmonic component. Nevertheless, we can place some bounds on β's range of variation.

7.4.5 Interpretation

Up to this point, we have set $\beta = \alpha = -72°$. However, this produced delays shorter than the acoustic propagation time from lips to microphone, i.e., $\tau_u < \tau_R$. This is not possible since if any part of the path is travelled at convection velocity, the delay will be increased. Therefore $\beta < \alpha$, i.e., β is more negative than α. Yet β has a lower bound, since otherwise we would observe phase wrapping during the pitch glides. (For the interval of a perfect fifth used here, the lower bound is -6π.) We thus have strong bounds on β: $-(3 \times 360)° < \beta < -72°$. In addition, we can compute the angle that would make the minimum τ_u just equal to the acoustic propagation of 2.9 ms: $\beta \lesssim -175°$.

The pitch glide data produced estimates of β that ranged from -120 to $-180°$, as presented in Table 7.1. The estimates so derived must be treated with caution for two reasons: they are based on one subject and only three glides, and the fitted lines are used to extrapolate an intercept value. Thus any variation in the glide itself will be magnified in the intercept estimate. By modifying the best fit lines to the pitch glide results, using one standard deviation to give the worst case gradients, we get a range of $-200° < \beta < -100°$. These weak bounds for the range of β, together with the stronger bounds given above, predict that β in Eq. 7.10 should lie within the range: $-200° \lesssim \beta \lesssim -175°$. Taking $\beta = -175°$ would effectively add 2.4 ms to the delays shown in the lower half of Figure 7.10.

7.4.6 Remarks

Inspection of the EGG signal during sustained phonation revealed a high correlation between the jitter and the shimmer of the glottal excitation, as has been widely reported. For subject PJ, the jitter and shimmer, sometimes collectively termed flutter, exhibited a typical natural frequency of ~ 7 Hz. Although f_0 glides were employed in the present study, salient time constants, such as these delays, could potentially be determined from the cross-correlation between the flutter (i.e., jitter or shimmer) and other derived signals (P_u, $\dot{P}_u$, etc.).

There were examples in the corpus of sustained fricatives where shimmer became so severe that voicing momentarily ceased. Our decomposition and modulation analysis was then applied to these regions to examine the effects of devoicing. The results are shown in Figure 7.10, beside the values for regular voicing, which are all on vertical grid lines. The repeatability of the phase values shows how the relation of the pulsing of the anharmonic component to the oscillation of the vocal folds is preserved across an interruption in phonation, despite the upset to the f_0 contour and consequently the decomposition algorithm.

Through an analysis of the relationship between place and the phase of the anharmonic signal envelope, a potential means of identifying unknown source locations has been unearthed, which also has implications for speech production studies of aspiration. However, the PSHF decomposition and demodulation analysis revealed a highly-variable phase for the epiglottal fricative /ʕ/. The phase of the modulation of the anharmonic component was $50 \pm 50°$, relative to the simultaneously-recorded EGG signal. This would imply a source location a short distance (a couple of centimetres) above the glottis, which would be consistent with sources distributed around the epiglottis.

The STP trajectories can be used as an objective means of defining consonantal duration, e.g., for [VCV] studies like Stevens et al. (1992). The phase profile appears to follow the expected path, but variability of phonation at voice onset and offset affect the decomposition, rendering the results inconclusive.

The heightened perceptual response of pulsed frication noise in voiced fricatives may be incorporated into normal speaker behaviour during development of the speech production faculty. It is possible that the articulatory configuration that maximises the pulsation, which occurs at specific Reynolds and Strouhal numbers, could become a natural articulatory target, since it tends to reduce the amount of effort required by the speaker to utter the voiced fricative. It is tempting to speculate further about the role of the observed phase differences in the categorical perception of voiced fricatives, particularly in opposition to aspiration noise, but we have found scant empirical evidence in the literature to support these claims and have performed no experiments of our own.

In summary, while it is clear that modulation of the anharmonic component varies with place, we can do no more than speculate that the acoustic-convective theory of sound production for the fricative component in voiced fricatives is the most likely, whose mechanism can be described as follows. A pulsed flow is emitted from the glottis into the vocal tract. Sound waves propagate down the vocal tract towards the constriction; at the constriction, the flow forms a jet, developing turbulence as it travels downstream. The temporal and spatial characteristics of the mixing flow are strongly influenced by the intersecting sound waves, inducing synchronous pulses of turbulence; the pulsed turbulence and entrained vortices convect downstream. When the jet encounters an obstacle (such as the teeth), a new source is generated that is pulsed at f_0 and efficiently radiates sound. The sound source at the obstacle excites the vocal tract; sound radiated from the lips propagates into the far field.

Assuming this to be the case, the increasing variance in Section 7.3.1 might be explained by three possible causes. First, the exact shape and location of the constriction may vary more for more posterior places, as the articulators become larger and are less finely controlled (e.g., tongue dorsum relative to tongue apex). Second, variations in convection velocity would make

a larger contribution for the more posterior fricatives where the vorticity has further to travel before reaching the obstacle. Third, the obstacle upon which the turbulence impinges is likely to extend further in the direction of flow, producing a more distributed source for constrictions nearer to the glottis.

7.5 Synthesis

Using what we have learnt about the timing relationship between the glottal pulses and the modulation of the fricative noise source, we present a modification to a simple synthesis procedure that more accurately reflects the relationship.

7.5.1 Source models

The [s, z] fricative pair was synthesised using the transfer functions calculated for [s] in Chapter 3. For voicing, a standard volume-velocity source was injected at the glottis, while the frication noise was positioned a short distance downstream from the constriction, as a pressure source at the teeth. The frication source was modulated by the glottal waveform for the voiced fricative with due regard to the timing of the pulses.

Our synthesis model was initially based on the assumption that the acoustic source and vocal-tract filter are independent. The frication source was convolved with the impulse response of the pressure VTTF from source to lips H^P_{QL} and the radiation characteristic, yielding a stationary noise signal, /s/. For its voiced counterpart /z/, the voiced source was generated using a cubic waveform, as in Klatt (1987), with an open quotient of 0.5 and fundamental frequency $f_0 \approx 130\,\mathrm{Hz}$. Filtering it by the volume-velocity VTTF from the glottis to the lips H^V_{GL} yielded the voiced component.

Many researchers have noted that the frication source appears to be modulated by voicing, e.g., Fant (1960), and the phase of the modulation has been shown to be perceptually significant (Hermes 1991; Skoglund and Kleijn 2000). Our analysis (above) confirmed that the noise component varied periodically according to fluctuations in the flow velocity at the constriction exit. Moreover, the modulation phase appeared to be governed by the convection time for the flow perturbation to travel from the constriction to the obstacle. Therefore, to synthesise the voiced fricative, the frication source $d(n)$ was modulated by the voice source $g(n)$ and the phase of its envelope delayed to match our empirical observations:

$$\hat{d}(n) = d(n)\, g(n - \tau)\,, \tag{7.11}$$

where the delay time τ was in the range 2.8–3.8 ms.

7.5.2 Results

As mentioned in Chapter 3, although we know that a critical parameter, the area of the constriction, has been over-estimated, we are interested in the performance of the entire chain, of which nearly every component has a novel aspect; it is more valid to compare results from varying source functions (e.g., from constant to two types of modulated noise) than to attempt to optimise the area function derived from dMRI. However, Narayanan et al. (1995) note only small differences between [s] and [z] for their subjects; the range of constriction areas is similar (from 0.12 to 0.25 cm^2 for [z]). Figure 7.11 (left) shows a portion of the synthesised voiced fricative /z/ with its constituent components. Alongside (right), there is a section of a sustained [z] recorded by speaker PJ (from the example in Fig. 7.1), which has been decomposed by the PSHF. Its harmonic and anharmonic components, act as estimates of the voiced and unvoiced signals, respectively. It is clear from the anharmonic signal that the noise has been modulated and that the peak amplitude is not synchronous with the glottal pulses, seen in the harmonic signal, although they share the same periodicity. In the synthetic components, the first formant dominates, yet the pulse-like excitation of the voiced components at glottal closure and the modulated envelope of the frication noise are characteristics echoed in the real signals.

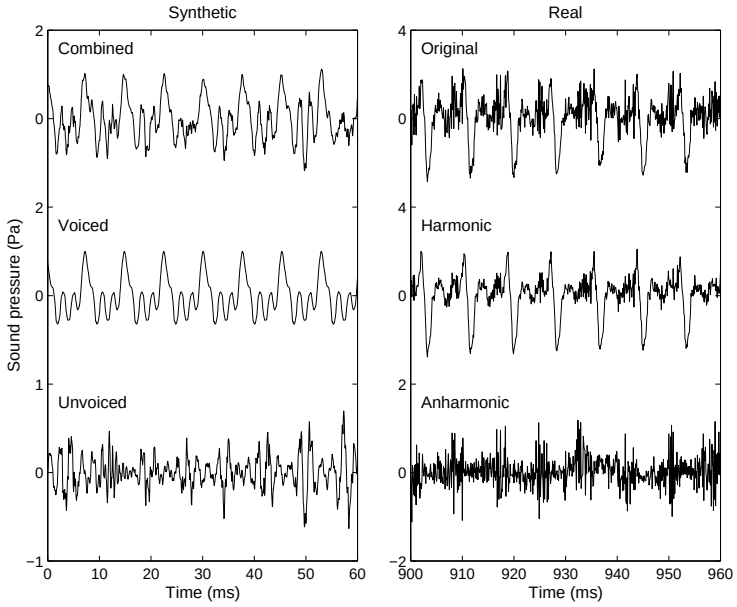


Figure 7.11: Synthetic (left) and real (right) signals for a sustained /z/ sound: (bottom, double amplitude scale) fricative component from a modulated noise source, (middle) voiced component, and (top) the combined signal.

Examples without the phase lag and with no modulation were also prepared for subjective assessment. These three synthetic examples of /z/ were all given the same harmonics-to-noise ratio (HNR = 6 dB): no modulation, modulation in-phase with the glottal waveform, and delayed modulation. Simple listening tests of the synthetic /z/ examples gave the following

subjective impressions. None of the examples sounded like a /z/, in part because the synthetic fricatives were presented without any transitions, and probably also because of the problems with the area function noted in Chapter 3. With the constant noise source, the noise seemed detached and the example unnatural. For the two modulated noise source examples, the sources were assimilated and did not give this detached impression, and the examples differed perceptibly from one another. The modulated noise source with the delayed phase relation, as found in real speech, sounded the most natural of the three.

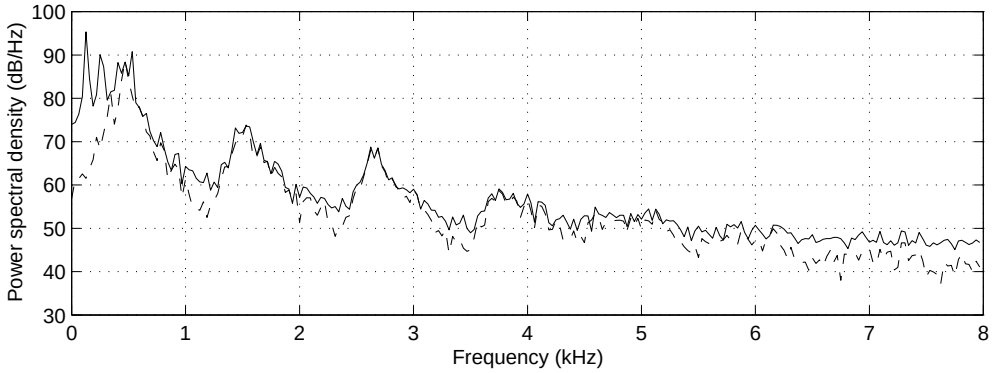


Figure 7.12: Power spectra of the synthetic voiced and unvoiced fricative pair: (solid) [z] and (dashed) [s].

The spectra of the voiced and unvoiced fricatives, shown in Figure 7.12, illustrate the effect of the voiced component. The voiceless spectrum has three prominent formant peaks (0.49, 1.54, 2.66 kHz), which give way to broader humps at higher frequencies (> 3 kHz), as seen in the VTTF (Fig. 3.8, top right). These formants are sometimes evident in measured spectra, but they usually peak in the 4–7 kHz band, and have a positive spectral tilt (see, e.g., Fig. 7.2). The presence of voicing has the effect of smoothing the spectrum at higher frequencies, as well as adding peaks at the first few harmonics of the fundamental frequency, f_0.

7.6 Conclusion

In this chapter, we have used the pitch-scaled harmonic filter (PSHF) as an quantitative technique for exploring source interactions. Voiced fricatives were decomposed into harmonic and anharmonic components. The amplitude of the components was represented by their short-time power (STP), which exhibited modulation at the fundamental frequency f_0. The relative phase of the modulation of the two components changes rapidly at a vowel-fricative transition, settling near an equilibrium that depends on the fricative's place of articulation. Fricatives at a range of places were recorded and analysed. The findings of this chapter support the suggestion that the aero-acoustic mechanism of fricative sound production is modified by voicing, due to the powerful effect of upstream acoustic disturbances as they intersect the

jet (Crow and Champagne 1971).

The PSHF algorithm was applied to give a plausible decomposition of the recorded utterance [ɑz:], successfully separating simultaneous parts of voiced and unvoiced speech. Inspecting the reconstructed time series, we observed the time-varying interaction of sources in the voiced fricative [z], manifested as pulsing of the unvoiced component, as had been noted in Chapter 6. Using the STP to approximate the signal envelopes, we derived an objective and quantitative method for measuring the magnitude and phase of the pulsation by complex demodulation. The phase difference between the modulation of the harmonic and anharmonic parts revealed two distinct states in the vowel-fricative transition. Referring the phase values to the EGG provided better fidelity in the modulation analysis and allowed us to attribute the change in state to the anharmonic component, which corresponded to a change in the unvoiced source location. The phase change decreased as the place of the constriction moved posteriorly, which was verified on a second subject (LJ).

A set of f_0 glide experiments showed that the phase, as a function of f_0, behaves almost entirely like a constant place-dependent delay. It is tempting to speculate further about the role of the observed phase differences in the categorical perception of voiced fricatives, particularly in opposition to aspiration noise, but we have found scant empirical evidence in the literature to support these claims. In perceptual tests on synthetic signals, Hermes (1991) found that the perception of noise bursts is affected by their phase relative to voicing; out-of-phase noise is distinguished from the voicing component, whereas synchronous bursts are assimilated.

In response, we attempted to synthesise a voiced fricative [z], modulating the frication noise with the appropriate time relation to voicing. By incorporating the delay observed between peaks in the glottal waveform and the envelope of the turbulence noise, we have created a source model that is more realistic from a physical and aero-acoustic perspective. Informal comparisons against the traditional in-phase modulation and no modulation cases favoured our model. In future, we plan to include transitions within phonemes (based on the adjacent dMRI frames), and between phonemes; these and improvements to other aspects of the synthesis will justify the use of more formal listening tests. We plan to include other speech sounds, e.g., stop consonants, and data for other subjects.

In short, we used the PSHF to decompose voiced fricatives into harmonic and anharmonic components. The different phase of the envelopes of these components led us to vary place and f_0 systematically for the purpose of determining the mechanism controlling the modulation. We have shown that a plausible explanation is that the acoustic signal generated at the glottis induces a structure in the jet emerging from the constriction, and thus alters the noise generated by the jet as it impinges on an obstacle. The second non-acoustic path that accounts for the variation of phase with place has not been incorporated into speech synthesis models until

recently (Sinder 1999). It would be instructive to ascertain whether Sinder's model predicts the phase changes we observed. It would also be useful to explore inter-subject variations and the robustness of phase changes to changes in f_0, effort and speaking style. Finally, the phase difference between harmonic and anharmonic components, which changes suddenly in the vowel-fricative transition, may well be perceptually important and should be investigated.[6]

[6]Further information can be found on the project website (Jackson 1998), including sound (`.wav`) files of the synthesised fricatives.

Chapter 8

Conclusion

8.1 Summary

This research project has two principal novel aspects: computing vocal-tract transfer functions (VTTFs) from a supraglottal source with the vocal-tract acoustics program VOAC, and the pitch-scaled harmonic filter (PSHF), a method for decomposing speech signals into harmonic and anharmonic components (representing the voiced and unvoiced parts of the signal, respectively). Although VOAC contains many features that make it more advanced than classical methods of one-dimensional acoustic modelling, the author's contribution has been to upgrade and extend the software to enhance its practical value so that it can predict VTTFs and perform basic speech synthesis, particularly for plosive, fricative and aspiration-noise sources. This was done by modifying the radiation impedance so that assumptions were valid over the frequencies important for speech, by adding the capability to calculate the VTTF from each source to radiated sound, and by providing the option of non-glottal source locations. Also the magnetic resonance imaging (MRI) data, upon which the present study depends, are an original source of dynamic, 3-D, vocal-tract information that supplied a realistic input to the acoustic model and were compared to speech analyses from the same subject.

The PSHF is an original technique for separating simultaneous components of mixed-source speech signals. Other techniques exist, but here our attention has been given to keeping the decomposed signals as faithful to the true voiced and unvoiced components as possible. As well as performance benefits of the PSHF over the alternatives in this sense, consideration of the end use of its outputs has been made explicit. Thus, signals that are destined for power spectral analysis or modelling are provided alongside signals for time domain analysis.

As a result of applying the PSHF to examples of a range of phonemes and in particular voiced fricatives, a relationship has been discovered between the timing of noise pulses in the anharmonic component and the place of articulation. In other words, the phase of modulation of the turbulence noise is a function of the constriction location. This mixed-source analysis

depends on the separation of simultaneous voiced and unvoiced components, and therefore demonstrates one way in which the development of these techniques can deliver new insights into the production of unvoiced speech sounds.

8.1.1 Acoustic modelling

The extension of VOAC to give VTTF predictions was first verified on experimental measurements from flow-duct models (Shadle 1985). These comparisons, which also tested the placement of the sound source downstream of the inlet to the tract, showed a match that was at least as good as, and arguably better than, that achieved with more classical acoustic modelling of the tube specimens. Furthermore, the losses associated with increased flow rate were mimicked automatically by VOAC. The results in Chapter 2 provide a preliminary indication of how the incorporation of flow as a factor in the calculation of the acoustic response of the vocal tract can be a benefit.

The interpretation of our MRI data to yield area (and hydraulic radius) functions for a number of different phonemes, including vowels and consonants, presented various challenges. There remain many questions about the details of converting sets of images to an accurate description of the vocal tract for use in a computer model. Difficulties of incorporating side branches into our one-dimensional model tended to exacerbate the issues surrounding their precise geometry. As discussed in Chapter 3, curvature of physiological structures such as the oral aperture can cause apparent misalignment between image slices that can devalue the resulting area functions, if not explicitly addressed. However, the main features of the VTTFs computed from dynamic MRI data were characteristic of their corresponding speech sounds and vowel formant frequencies agreed with expected values. Finally, the VTTFs were used to generate speech-like signals corresponding to each of the phonemes /p, ɑ, s, i/.

8.1.2 Speech analysis

Having formed a working definition of aspiration at the outset of this report, namely "flow-induced turbulence noise that is not frication", we investigated a number of standard analysis techniques for the purposes of extracting its characteristics from a series of speech recordings. By aligning repeated tokens of unvoiced plosives and then ensemble averaging, we were able to enhance their common features, which yielded a number of specific findings. An estimate of the source location was calculated from spectral troughs, corresponding to anti-resonances in the VTTF from the rear-tract resonances, which was illustrated for the bilabial plosive /p/. Differences from the effects of place were evident in the patterns of both resonances (formants) and anti-resonances (zeros) in the averaged spectra.

The requirement to study the nature of unvoiced sounds in all speech, with and with-

out voicing, has led to the development of a signal processing technique that can practically separate harmonic and anharmonic contributions in the speech signal, corresponding to the voiced and unvoiced components, respectively. Some other methods have been considered, particularly the PAPD method in Appendix D, which come from related speech applications (e.g., voice quality, synthesis or enhancement) and similar decomposition objectives, leading to further developments.

The PSHF analysis technique has been developed for decomposing mixed-source speech signals, and addresses the twin goals of reconstructing signals for subsequent time-series and power-spectrum analysis. Based on a pitch-scaled separation in the frequency domain, the PSHF estimates the voiced and unvoiced components, using only the speech signal. The PSHF was designed to be robust to the sorts of variation and perturbation typically observed in speech, while retaining most of the performance obtained by a maximum likelihood approach. Tests on synthetic speech demonstrated the PSHF's ability to reconstruct time series corrupted by jitter, shimmer and additive noise, and implied improvements to the signal-to-error ratio (SER) of $\eta_u \approx 14\,\mathrm{dB}$ to the anharmonic part in normal speech conditions (decreasing with increased corruption).

Processing real speech examples resulted in convincing decompositions, extracting and revealing features particular to the individual components. In agreement with the predictions in Chapter 5 at various values of f_0, the PSHF was shown to be robust for both male and female speakers. Results were presented for the nonsense word [p$^{\mathrm{h}}$ɑzɑ] in Chapter 6. The algorithm performed well in steady conditions, revealing features of the unvoiced sounds that were previously masked by voicing, but suffered degradation from jitter, shimmer and rapid transients. Earlier (in Chapter 5), the tests showed in a more precise manner how the respective performance degradations tallied. Local measurements of the perturbation of the recorded speech signal were then used to predict the fidelity of the voiced and unvoiced estimates. Analysis of speech with various voice qualities showed that breathiness affected more than just the proportion of noise in the speech signal: the shape of the glottal excitation, the degree of variability in voicing as expressed by jitter and shimmer metrics, the damping of resonances and the overall shape of the noise spectrum.

The decomposition of the speech signal into estimates of the voiced and unvoiced components in the frequency domain, and their reconstruction into time-series signals enables parallel analyses to be performed on the components. Using standard techniques to extract the features, differences can give information about salient contrasts. Moreover, by devising ways of exploiting the synchrony of the signals, new methods of analysing mixed-source signals can be created. One example, which explored the interaction of voicing and the production of frication noise, was described in Chapter 7. Pulsing of the noise component was observed in a voiced

fricative [z], which was analysed by complex demodulation of the signal envelope to reveal a sort of time-varying source interaction. What was taken to be breath noise during vowels appeared to pulse in-phase with the periodic oscillation of the voiced component, whereas in [z], the pulses showed a very different phase relation. The timing of the pulsation, represented by the phase of the anharmonic modulation coefficient, showed a step change during a vowel-fricative transition [ɑz], corresponding to the change in location of the sound source within the vocal tract. A study of other fricatives demonstrated the relationship between phase and place, and f_0 glides confirmed that the main cause was a place-dependent delay, whose origin we endeavoured to explain. An attempt was made to synthesise primitive fricative examples to illustrate the effect of the phase change. These speech-like signals confirmed reports of the perceptual significance of the phase relation in simple listening tests.

8.2 Findings

Above and beyond the development of acoustic modelling and speech analysis tools summarised in the previous section, there are various findings from this research that are of specific relevance to certain kinds of unvoiced sound. Our goal was to make enhancements to a generalised model of sound production. While many of our analyses merely confirmed either the findings of others or our suspicions, substantial discoveries were made that imply significant changes with respect to existing models.

8.2.1 Fricatives

Using the aforementioned PSHF decomposition and parallel analysis techniques from Chapter 7, it was found that the pulsing of the frication noise during voicing was dependent on the location of the constriction in the supraglottal tract. In fact, the timing of pulses appeared to be governed by fluid convection along the tract, downstream of the constriction. This result is consistent with earlier observations of the interaction of sound waves and flow turbulence, and firmly suggests that a solely acoustic representation of the sound generation provides an inadequate description of the real process. The noise-modulation behaviour has been shown for a full range of voiced fricatives, but it may also be relevant to other mixed-source sounds, such as voiced plosives and breathy vowels.

By comparison, our spectral analysis of voiced and unvoiced fricatives in Chapter 6 showed that the frication noise has consistent characteristics in the two cases. This analysis was performed using ensembles of fricative tokens uttered in identical contexts by a single speaker. The contribution from the voicing source was removed by the PSHF and the remaining anharmonic components were averaged, as were the unvoiced fricative tokens. All the significant features of

the mean power spectra matched, suggesting that the spectral characteristics of the frication source and the source-to-far field transfer function were not notably affected by voicing.

The results of the vocal-tract transfer function predictions for /s/, however, were somewhat disappointing. The frequencies of resonances and anti-resonances seemed to have been predicted with a fair degree of accuracy, but the extent of the losses was widely underestimated, resulting in bandwidths that were too narrow. Also, the gross spectral features and overall spectral tilt did not reflect those observed from corresponding speech recordings. These factors raise questions about the accuracy of the source functions (see e.g., Narayanan et al. 1995) and of the area functions, especially near the constriction. The former may be addressed by supplementing the MRI data with other sources of data specifically referring to the intra-oral constriction.

8.2.2 Plosives

The VTTF calculated for the plosive /p/ gave a good match of the measured peaks and troughs in the spectrum up to 7 kHz, although greater losses were needed in the low-frequency region. In the spectra of the ensemble-averaged measurements, many of the burst features were surprisingly clear. There were distinct spectral troughs which, since their frequency was approximately linear with distance from the glottis, changed slightly for different places of articulation. More significantly, perhaps, the relative amplitudes of the formants were radically altered by changing the place. These spectra were found to resemble those of the co-located voiceless fricative. Analysis of the time-varying sound sequence following the release of an aligned set of bursts illustrated the changing aural scene that was generated. The burst and onset of voicing events were easily identified as the beginning and end of the sequence, but although there were notable variations, the fricative and aspiration stages were less simple to demarcate.

8.2.3 Aspiration noise

The examples of aspiration that were examined covered a range of contexts, being either embedded within some other dominant sound, as in a breathy vowel, or a stage estimated from a snapshot after plosive release. In all cases, the spectral envelope was similar to that of a vowel for the same vocal-tract configuration, but with a very different spectral tilt which, unlike a vowel, would typically be positive over the first 8 kHz or so. Accordingly, the first formant would tend to be very weak. The source was broad-band noise by nature, so that the resultant spectrum was usually flatter than either the positive-tilt burst stripe or the negative-tilt frication spectrum.

Decomposition of vowels suggested that aspiration noise might be modulated, just like

frication noise in the presence of voicing. However, because of the high harmonics-to-noise ratios involved, these results are inconclusive.

8.3 Future work

The future directions of potentially fruitful research fall into two main categories: (i) enhancements to the tools and techniques used in this project (in Sections 8.3.1–3), and (ii) extensions to the scope of their application to speech (Sections 8.3.4 onwards).

8.3.1 The VOAC program

While the results of the VOAC program's calculations are insightful and worthy (as this thesis demonstrates), it achieves them in an awkward manner. A future objective is to make the program suite publically available for research purposes, yet its routines are currently difficult to understand and slow in execution. Its translation into Matlab has made it much easier to debug, develop and maintain for an experienced user, but many improvements are needed to encourage more widespread application to speech. Many internal loops could be converted to vector or matrix operations, with the questionable merit of brevity, yet much more work would be required to make the structure inside the main subroutines transparent. There exist many opportunities for increased modularisation as the program stands, but by reorganizing the encryption of geometry functions, more dramatic performance gains can be won.

Since translation of VOAC v4.0 from Fortran to Matlab, a newer version (v4.5) of the Fortran code has been made available. Its structure is much simpler than the older version, since it does not break the area function down into element types, but considers the transfer at each sub-element boundary in turn (see Fig. 2.1). At a boundary, it decides whether the area gradient $\delta S/\delta x$ is gradual (slow) or abrupt (fast) by comparison of the area change over the length of the sub-element against a threshold. Transfer in slow changes is computed as for cylindrical wavefronts (cf. Type 2), and fast changes as for an expansion or contraction, similar to Types 1 and 4 before. Also incorporated is the facility to add simple side branches, as in the previous Type 4 element. Thus, it can be seen that the only option lost in the newer structure is the Type 3 conical element[1] (since Type 5 is ostensibly included in Type 2). The re-design results in the conversion from real area function data to input files being simpler, and provides a more intuitive representation of the geometrical information. In addition, its implementation in program code is shorter than the older version, and so its maintenance would be less onerous.

The notion of simpler primitives from which to construct the area functions is an obvious path to take. It would abbreviate and modularise whole sections of code, both desirable

[1] In any case, the CONE element (Type 3) is not functioning currently (see Appendix B.1).

programming qualities since they simplify the maintenance task greatly in comparison to that of the current code which contains composite element types. Routines for reading, writing and displaying geometry functions would also benefit as a consequence, and new types of geometry could more easily be incorporated. In particular, the option to include side branches at angles other than 0° or 180° to the medial axis might be considered (Dang et al. 1997).

It would be intellectually satisfying to extend the derivations of the mathematical formulae given in Appendix A so as to explain all aspects of the program code. More practically, the expression for the radiation impedance, which is based on a piston in an infinite baffle at present, may not be the most appropriate. The effects of replacing this with a more accurate expression should be investigated, such as that for a piston in a sphere. Another practical measure concerns the depiction of non-terminal sources, like those attributed to frication near a supraglottal constriction. Currently, the program requires a new geometry function to be generated for each source location, in addition to that of the complete vocal tract. Clearly, a more tractable solution can be implemented to avoid such replication of data, notwithstanding the need for separate VTTFs.

The longer-term goal of using VOAC to generate high quality synthetic speech would realistically depend on finding a straightforward means of transmitting the VTTF information into a state-of-the-art articulatory synthesis system. Using dMRI data, for example, it could be transformed into a complete hybrid synthesis system (Sondhi and Schroeter 1987). For use as an analysis tool, its output may need to be interpreted to give parameters that can be compared directly with values obtained from real speech. For instance, the frequencies of resonances can be compared to the output of a formant tracker. However, for other features, it may be necessary to develop our methods of analysis to yield a suitable parameter to describe the desired feature.

8.3.2 Speech analysis

For acoustic sources not at the glottis, the frequencies of anti-resonances are a characteristic spectral feature. Ways of reliably extracting this information from speech recordings have not yet been devised, although algorithms exist for deriving ARMA system models, e.g., Akaike, Yule-Walker (Yegnanarayana 1981; Childers 2000), whose transfer functions contain both poles and zeros. To increase the robustness of such estimation procedures, it may be advantageous to perform cepstral smoothing of the power spectra, as pre-processing. Having extracted such features, it may be possible to identify the source locations by tracking the zeros obtained from processed speech and computing the rear-tract length, as in Chapter 4.

One form of the analysis that has not yet been fully exploited in the context of mixed-source speech, is pitch-synchronous analysis (Pinson 1963; Shadle 1995a; Yegnanarayana and Veldhuis

1998). Applied to the outputs of the PSHF, this analysis could provide more precise details of the variation of both the filter (Rothenberg 1981) and the source during the glottal cycle. These new techniques for analysing mixed-source speech open up the possibility of new kinds of phonetic study, which might use the short-term HNR, say, as an objective speech segmentation aid or means of estimating phoneme durations.

Finally, increased confidence in the quality of VTTF predictions from our acoustic model would allow for analysis of complex sounds using VOAC. For instance, some of the more complex sounds that we have examined, stop consonants and mixed-source speech, could benefit from an approach that uses a priori knowledge of the nature of speech sounds to find the most likely interpretation of the observations. This approach could be implemented as a procedure for minimising the error between a hypothesised VTTF and some form of the observed speech signal. It might, for example, be a least-squares fitting of a VTTF sequence multiplied by a source spectrum to the short-term magnitude spectrum, where the sequence was generated for a set of potential constriction locations. Alternatively, the frequency space might be perceptually weighted, or the comparison could be made between mel-frequency cepstral coefficients. Methods of this kind have been used in attempts at estimating the area function from the acoustic signal (Heinz and Stevens 1961; Shirai and Masaki 1983; Badin et al. 1995; Story and Titze 1998a).

8.3.3 Mixed-source decomposition

Several alternatives have been published as methods of decomposing speech signals into estimates of the voiced and unvoiced components, some of which were described and tested in Chapter 5. For any particular task, the most appropriate method must be selected, correctly implemented and applied to the target data. A methodology has been developed as the first step in defining a protocol for benchmarking the decomposition performance of alternative methods (in Chapter 5 and Appendix D, and in d'Alessandro et al. 1998), and this work should be extended to all the popular methods. This would allow a meaningful comparison to be made, which would not only facilitate the selection of the best method for a certain task but would give a quantitative view on future developments. Thus, augmentation of the PSHF, by assimilation of an alternative's peripheral processing of modification of one or other component of the algorithm, could be assessed in a more rigorous way, indeed as could any competing method.

The performance of the PSHF could be improved by the following means. Trying to whiten the speech signal before decomposition is a way of minimising the impact of abrupt operations in the frequency domain. The spectral leakage of the skirts of a harmonic obtained from a windowed signal will have a lesser effect on its neighbouring harmonics if their amplitudes

are evenly balanced beforehand. However, as mentioned in Appendix D, when the mixed sources have very different spectral tilts, this can be counterproductive. Yet there may be situations when this approach can be of benefit. Tailoring the whitening process to each source in turn might improve results with respect to the raw speech, with individual estimates of the spectral envelope for each source as a by-product. Other advances may be made by trying to improve the way vocal-fold oscillation is modelled, which could be some form of Kalman filter, possibly treating it as a non-linear oscillator. It is helpful to try to incorporate whatever prior knowledge we have about the nature of the signals. So far we have attempted to do so using crude assumptions of harmonicity and spectral flatness for deterministic and stochastic components, respectively, yet Bayes' theorem provides us with a framework for incorporating many other attributes of the speech signals.

8.3.4 Extension of speech corpus

Ultimately, the real benefits of progress in speech decomposition technology are to be measured by what it enables us to discover about speech production. Therefore, studies of the properties of the output signals are the key to how speech characteristics vary with phoneme, context, sex, f_0, open quotient, mode of phonation, speech rate, and a host of other parameters. Other opportunities for investigation include applying pitch-synchronous analysis to the existing corpora, that have simultaneous EGG traces (*C4–6* in Chapter 4), a wider study of the noise-modulation phase, and evaluation of the perceptual effects of modulation. The wider study might encompass static tests across a larger subject pool, that comprise recordings of sustained fricatives and even f_0-glides, and dynamic tests where the phoneme in question is set in a carrier phrase. Evaluation of the perceptual significance of modulation is not fully understood, despite the pioneering work of Hermes (1991) and Strope and Alwan (1998). The phase difference of the modulation is known to affect the assimilation of mixed-source components, but does it influence the categorical perception of fricative and aspirative sounds? Alterations to the synthesis of these sounds in a high-quality, natural speech synthesiser would enable formal listening tests to be performed in a truly representative manner.

8.3.5 Interpretation of images

While the results of this project have highlighted the difficulty of gaining accurate geometrical descriptions from a series of magnetic resonance images, the data derived showed the potential for creating speech from them. Moreover, the fact that these images had such fine time resolution leads the way to many avenues of future research. The dynamics extracted from the vocal-tract outlines and their area functions can be compared directly with other forms of articulatory data, either from measurement or from a model. Thus, the dynamic MRI information

offers a route to a better understanding of articulatory dynamics, which would be expected to pay dividends in an articulatory-based synthesis system. Once individual phones have been examined, the transitions within phonemes and between phonemes can be investigated, using adjacent dMRI frames. However, further work is needed to improve the conversion of MRI frames to geometry functions, following the questionable renditions in Chapter 3.

8.3.6 Physical flow models

To validate the acoustic model, it is best to test the solutions hypothesised by VOAC against experimental flow-duct data. Once satisfied with the accuracy of predictions of the filter characteristic of the flow-duct transfer function, the apparatus can be used to study aspects of the source. One experiment of particular relevance to the present study would be to examine modulation behaviour of the flow noise. Using the methods developed in Chapters 5 and 7, precise measurements of the phase could be established under the controlled conditions of a physical test rig. Hence, the vortex source location and convection velocity could be deduced with a reasonable degree of accuracy for voiced-fricative type configurations.

8.4 Coda

In this thesis, a flow-duct acoustic modelling tool has been further developed, and used to predict the acoustic response of vocal-tract configurations measured by dMRI. Analysis of plosive releases has identified a number of properties characteristic of the place of occlusion and of sounds following the burst. In addition, a technique for separating the voiced and unvoiced components of a speech utterance has been proposed, which has been shown to provide good performance over a wide range of conditions, through tests on synthetic speech-like signals. Accurate and convincing decompositions of real speech were achieved, with a fair degree of robustness, for example, extracting aspiration noise from recorded vowels. Applying the technique to voiced fricatives, a timing relationship between voicing and the generation of turbulence noise was discovered that was at odds with conventional models of speech production. Although corresponding modifications were made to the noise-source model, further work is needed to synthesise these sounds naturally.

Appendices

Appendix A

Acoustic transfer equations

A.1 Fundamental relations

The following formulae show the derivation of the aero-acoustic equations used by the program VOAC. Expressions for simple geometric primitives are elaborated, starting from the basic physical and thermodynamic relations: continuity of mass, conservation of momentum and conservation of energy. The final equations are derived through linearisation with respect to the acoustic partial pressures, while retaining terms that depend on the flow. The appendix begins with a statement of physical constants in Table A.1, and standard acoustic expressions. A control volume is defined, for which the laws of conservation are derived, each in turn. Then, transfer equations are obtained for a contraction, an expansion and a side branch (no flow) for direct comparison with VOAC's pseudocode, which is listed in Appendix B. Finally, expressions for the radiation impedance under assumptions of alternative boundary conditions are given.

A.1.1 Acoustic equations (no flow)

For comparison, statements of the standard acoustic equations for continuity of mass,

$$\frac{\partial \rho}{\partial t} + \rho_0 \frac{\partial u}{\partial \mathbf{x}} = 0\,, \tag{A.1}$$

and conservation of momentum,

$$\rho_0 \frac{\partial u}{\partial t} + \frac{\partial p}{\partial \mathbf{x}} = 0 \tag{A.2}$$

are included here, which combine to give the wave equation:

$$\left(\frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} - \nabla^2 \right) \phi = 0\,. \tag{A.3}$$

Parameter	Wet, warm	Dry, warm	Dry, ambient	Dry, zero	Units
temperature T	310 (37)	310 (37)	293 (20)	273 (0)	K (° C)
relative humidity RH	100	0	0	0	%
speed of sound c_0	[e]359	[e]353	343	[d]331	m/s
density ρ_0	[e]1.098	[e]1.139	[a]1.205	[d]1.29	kg/m^3
density-speed product $\rho_0 c_0$	394	402	413	-	kg/m^2 s
gas constant R	[b,e]297.7	[e]287.0	[b]287.0	*287.0	J/kg
specific heat c_p (constant pressure)	-	-	[b]1.01×10^3	-	J/kg K
ratio of specific heats γ	[e]1.396	[b,e]1.400	[b]1.400	[d]1.400	-
absolute viscosity μ	*1.89×10^{-5}	[e]1.89×10^{-5}	[a]1.81×10^{-5}	[d]1.71×10^{-5}	kg/m s
kinematic viscosity ν	*1.66×10^{-5}	[e]1.66×10^{-5}	[a]1.50×10^{-5}	-	m^2/s
coefficient of heat conduction λ	-	-	-	[c]0.024	J/m s K

Table A.1: Thermodynamic constants for air at atmospheric pressure, $p_0 = 1.013 \times 10^5$ Pa. Key to sources: [a]Beranek (1954), [b]Table 2, p. 2 (Haywood 1968), [c]Table 22, p. 33 (Haywood 1968), [d]Table XIV, p. 905 (Morse and Ingard 1968), [e]Appendix, p. 63 (Hardcastle and Laver 1997). *Values inferred from those adjacent.

A.1.2 Isentropic and adiabatic processes

For perfect gases undergoing an isentropic process, pressure p' and density ρ' (the reciprocal of specific volume) are related by the equation

$$p'/\rho'^{\gamma} = m\,, \tag{A.4}$$

where m is a constant. We define the total pressure p' and density ρ', as the sum of a small perturbation, p or ρ respectively, and the time-averaged (mean) values, p_0 or ρ_0 (denoted by the subscript zero) such that,

$$p' \doteq p_0 + p\,, \text{ and} \tag{A.5}$$

$$\rho' \doteq \rho_0 + \rho\,. \tag{A.6}$$

Using this nomenclature, differentiating and then substituting back in,

$$\begin{aligned} p &= m\gamma {\rho_0}^{\gamma-1}\,\rho \\ &= \frac{p_0}{\rho_0^{\gamma}}\gamma {\rho_0}^{\gamma-1}\,\rho \\ &= \gamma\frac{p_0}{\rho_0}\,\rho\,. \end{aligned} \tag{A.7}$$

Hence, using $RT_0 = \frac{p_0}{\rho_0}$ in Eq. A.7, the fluctuating quantities can be related by:

$$p = \gamma\frac{p_0}{\rho_0}\,\rho$$

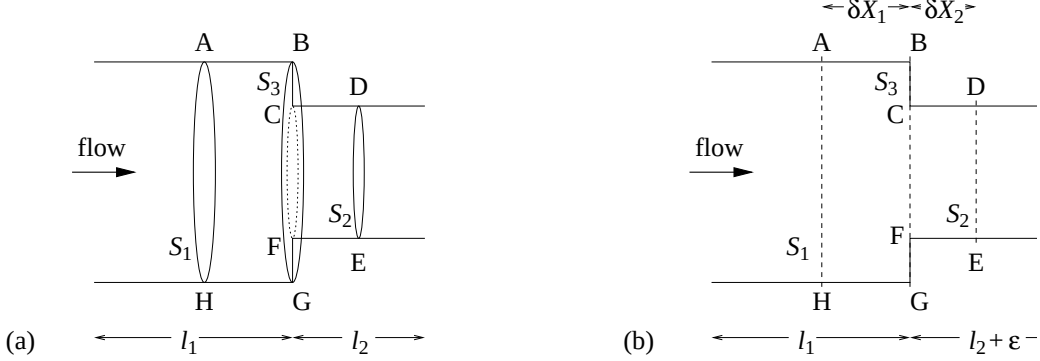


Figure A.1: Diagram of the control volume ABCDEFGHA for a contraction, indicating (a) the boundary planes S_1, S_2 and S_3, which are, respectively, upstream of, downstream of, and at the area discontinuity; and (b) the geometry modified by the end-correction ϵ, as used by VOAC.

$$= \gamma R T_0 \rho$$

$$= c_0^2 \rho \,, \tag{A.8}$$

where the speed of sound is defined as:

$$c_0 = \sqrt{\gamma R T_0} \,. \tag{A.9}$$

The flow velocity can also be written as the sum of the mean flow, i.e., the net flow u_0, and a fluctuating component, which is the acoustic velocity u:

$$u' \doteq u_0 + u \,. \tag{A.10}$$

The acoustic impedance z, analogous to its electrical correlate, can be written as the ratio of the acoustic pressure and velocity at any point:

$$z \doteq \frac{p}{u} \,, \tag{A.11}$$

where pressure p is equivalent to potential difference (voltage) and velocity u to current. For plane wave propagation in free-space, the impedance is $z = \rho_0 c_0$.

A.1.3 The control volume

For computing the transfer of acoustic pressure across an abrupt change in area, we consider a control volume ABCDEFGHA that surrounds the discontinuity. The control volume is limited by the duct walls and three planes: the two ends AH and DE, and the junction BCFG, which have cross-sectional areas S_1, S_2 and the difference S_3, respectively. They are shown for a contraction in Figure A.1a, in which case $S_1 = S_2 + S_3$ whereas, for an expansion, $S_1 = S_2 - S_3$. The way that the acoustic end-effects of the discontinuity, which are the result of cross-mode

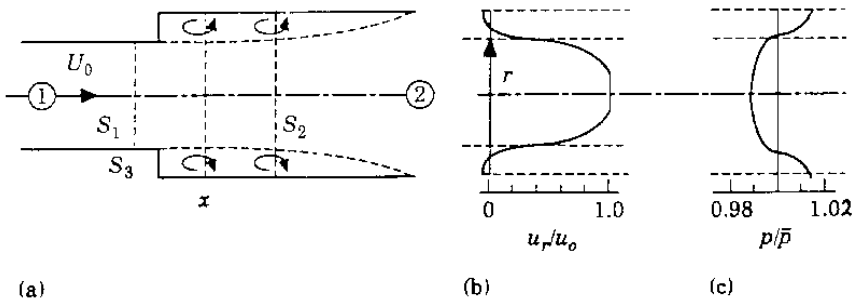


Figure A.2: Diagram of an expansion (reprinted from Davies 1988, Fig. 3, p. 100): indicating (a) the duct geometry, the control volume planes and the region of net flow; and the profiles at S_2 of (b) the mean velocity; and (c) the fluctuating pressure.

matching of the boundary conditions, are incorporated is illustrated in Figure A.1b (see Section 2.3.5). They are modelled by inclusion of an end-correction factor, which is empirically defined as (Eq. 4.10 in Davies 1988, p. 104):

$$\epsilon = \begin{cases} r_2\kappa \left[1 - \exp\left(\frac{1-\sqrt{S_1/S_2}}{a}\right)\right] & \text{for } S_1 > S_2 \\ r_1\kappa \left[1 - \exp\left(\frac{1-\sqrt{S_2/S_1}}{a}\right)\right] & \text{for } S_1 < S_2 \end{cases} \tag{A.12}$$

where $\kappa = 0.63$, $a = 1.5$ and the hydraulic radius $r = 2S/l_\pi$ for perimeter l_π.

The positive direction for the net flow is from left to right, that is, from region 1 to region 2, by our convention. In the presence of flow, an abrupt increase in area causes flow separation and the formation of a jet at the expansion. For such cases, the downstream plane S_2 transects the effluent jet, before turbulent mixing has fully developed to a uniform flow profile. Figure A.2 shows the path of the effluent jet and its mean velocity and acoustic pressure profiles for the cross-section at S_2, where mass flux, momentum and enthalpy are evaluated for the control volume. As the control volume boundaries are brought nearer together ($\delta X_1 \to 0$ and $\delta X_2 \to 0$), such that S_1 and S_2 are just upstream and downstream of the area change respectively, the effects of the side-walls become negligible and we can equate mass momentum and energy either side to resolve the reflection and absorption of plane acoustic waves, travelling in a duct.

A.2 Continuity of mass

The linearised equation for continuity of mass (Eq. 4.1 in Davies 1988, p. 101) across the control volume boundary is defined for an expansion by:

$$\int_{S_1} (\rho_0 u + u_0 \rho)_1 \, dS_1 + \int_{S_3} (\rho_0 u + u_0 \rho)_3 \, dS_3 - \int_{S_2} (\rho_0 u + u_0 \rho)_2 \, dS_2 = 0\,, \tag{A.13}$$

where the subscript zero refers to the time-averaged (steady-state) quantity and non-zero subscripts refer to the planes 1, 2 and 3 respectively, as in Figure A.2. For a contraction, the

integral over S_3 would change sign, but since there is no mass flux across the duct wall, the integral can be eliminated, being equal to zero, to make the equations for a contraction and an expansion identical.

We will consider the fluctuating quantities to be comprised of an infinite Fourier series of complex sinusoidal components: $p(\omega)\exp j\omega t$, $\rho(\omega)\exp j\omega t$ and $u(\omega)\exp j\omega t$. For plane-wave propagation, the pressure $p(\omega)$, has two contributions: one from the positive-travelling wave $p^+(\omega)$, and the other from the negative-travelling wave $p^-(\omega)$. The positive direction is defined as being the same as that of the flow, and vice-versa. This convention can also be applied to $\rho(\omega)$ and $u(\omega)$ to describe the effect of the positive- and negative-travelling components. For clarity, the frequency dependence has been omitted, but is implied in all expressions containing the terms p^+ and p^-. We can write the acoustic pressure in terms of the positive- and negative-travelling waves thus: $p = p^+ + p^-$. Similarly, the acoustic velocity can be written $u = (p^+ - p^-)/\rho_0 c_0$. The density is normally $\rho = p/c_0^2$ (Eq. A.8), but, to account for the losses associated with the turbulent mixing in shear layer after an expansion, we then include the loss term δ, according to Eq. 2.5 in Davies (1988, p. 94): $\rho = (p^+ + p^- + \delta)/c_0^2$.

A.2.1 Contraction

If we take an area contraction inside the control volume, and consider the flow to be uniform and isentropic on both sides of the discontinuity (i.e., $M = M_1 = u_0/c_0$, and $M_2 = u_0 S_1/c_0 S_2 = MS_1/S_2$, and zero mixing losses $\delta = 0$), the expression for mass flux becomes:

$$S_1\left(\rho_0\frac{p_1^+ - p_1^-}{\rho_0 c_0} + M\frac{p_1^+ + p_1^-}{c_0}\right) = S_2\left(\rho_0\frac{p_2^+ - p_2^-}{\rho_0 c_0} + \frac{S_1}{S_2}M\frac{p_2^+ + p_2^-}{c_0}\right) . \tag{A.14}$$

Multiplying through by c_0 and collecting the pressure terms, we have:

$$S_1\left[(1+M)\,p_1^+ - (1-M)\,p_1^-\right] = S_2\left[\left(1+\frac{S_1}{S_2}M\right)p_2^+ - \left(1-\frac{S_1}{S_2}M\right)p_2^-\right] . \tag{A.15}$$

A.2.2 Expansion

Re-evaluating the integrals in Eq. A.13, remembering that the net flow u_0 applies only to an area equal to S_1 at the S_2-plane (section D-E in Fig. A.1), we have:

$$\begin{aligned} S_1\left(\rho_0 u + u_0\rho\right)_1 &= S_2\left(\rho_0 u\right)_2 + \int_{S_2}\left(u_0\rho\right)_2\, dS_2 \\ &= S_2\left(\rho_0 u\right)_2 + S_1\left(u_0\rho\right)_2 . \end{aligned} \tag{A.16}$$

Recalling $M = u_0/c_0$ and substituting, Eq. A.16 becomes:

$$S_1\left(\rho_0\frac{p_1^+ - p_1^-}{\rho_0 c_0} + M\frac{p_1^+ + p_1^-}{c_0}\right) = S_2\left(\rho_0\frac{p_2^+ - p_2^-}{\rho_0 c_0} + \frac{S_1}{S_2}M\frac{p_2^+ + p_2^- + \delta}{c_0}\right) . \tag{A.17}$$

Multiplying through by c_0 and collecting the pressure terms, as before, we have (Eq. 4.2, Davies 1988, p. 102):

$$S_1\left[(1+M)\,p_1^+ - (1-M)\,p_1^-\right] = S_2\left[\left(1+\frac{S_1}{S_2}M\right)p_2^+ - \left(1-\frac{S_1}{S_2}M\right)p_2^- + \frac{S_1}{S_2}M\delta\right] \quad \text{(A.18)}$$

which is identical to Eq. A.15, with the addition of the $M\delta S_1/S_2$ term.

A.2.3 No flow

In the case of zero net flow velocity ($M = 0$, $\delta = 0$), Eq. A.18 reduces to (Eq. 4.3, Davies 1988, p. 102):

$$S_1\left(p_1^+ - p_1^-\right) = S_2\left(p_2^+ - p_2^-\right)\,. \quad \text{(A.19)}$$

A.3 Conservation of momentum

Equating the resultant axial force due to pressure to the net momentum flux, the conservation of momentum across the plane of transfer is defined for an expansion by:

$$\begin{aligned}&\textstyle\int_{S_1} p_1'\,dS_1 + \int_{S_3} p_3'\,dS_3 - \int_{S_2} p_2'\,dS_2 \\ &\qquad = \textstyle\int_{S_2} \rho_2'(u_2')^2\,dS_2 - \int_{S_1} \rho_1'(u_1')^2\,dS_1 - \int_{S_3} \rho_3'(u_3')^2\,dS_3\,.\end{aligned} \quad \text{(A.20)}$$

As before, the integrals over s_3 change sign for a contraction. The integrands on the right-hand side can be expanded and then linearised by deleting the second order terms and higher:

$$\begin{aligned}\rho'(u')^2 &= (\rho_0+\rho)\left(u_0^2 + 2uu_0 + u^2\right) \\ &= \rho_0u_0^2 + 2\rho_0u_0u + \rho_0u^2 + \rho u_0^2 + 2u_0\rho u + \rho u^2 \\ &\approx \rho_0u_0^2 + 2\rho_0u_0u + \rho u_0^2\,,\end{aligned} \quad \text{(A.21)}$$

where non-zero subscripts refer to fluctuating quantities in the respective numbered region. Noting that $u_3' = 0$, subtracting the steady-state (time-averaged) values and using the above approximation, Eq. A.20 linearises to:

$$\begin{aligned}&\int_{S_1} p_1\,dS_1 + \int_{S_3} p_3\,dS_3 - \int_{S_2} p_2\,dS_2 \\ &\qquad = \int_{S_2}\left(\rho u_0^2 + 2\rho_0u_0u\right)_2\,dS_2 - \int_{S_1}\left(\rho u_0^2 + 2\rho_0u_0u\right)_1\,dS_1\,.\end{aligned} \quad \text{(A.22)}$$

A.3.1 Contraction

Observing the sign change of the S_3 term, evaluation of the integrals in Eq. A.22 for a uniform flow across S_2 with no losses ($\delta = 0$) yields:

$$S_1p_1 - S_3p_3 - S_2p_2 = S_2\left(\rho u_0^2 + 2\rho_0u_0u\right)_2 - S_1\left(\rho u_0^2 + 2\rho_0u_0u\right)_1\,, \quad \text{(A.23)}$$

which, since $p_3 = p_2$, can be re-written,

$$S_1 p_1 - (S_3 + S_2)\, p_2 = S_2 \left[\left(\frac{S_1}{S_2} M \right)^2 p_2 + 2 \frac{S_1}{S_2} M u_2 \right] - S_1 \left[M^2 p_1 + 2 M u_1 \right] . \tag{A.24}$$

Now, recalling that $S_1 = S_2 + S_3$, usual substitution gives,

$$\begin{aligned} & S_1 \left(p_1^+ + p_1^- \right) - S_1 \left(p_2^+ + p_2^- \right) \\ & \quad = S_1 \left[\tfrac{S_1}{S_2} M^2 \left(p_2^+ + p_2^- \right) + 2M \left(p_2^+ - p_2^- \right) \right] - S_1 \left[M^2 \left(p_1^+ + p_1^- \right) + 2M \left(p_1^+ - p_1^- \right) \right] \end{aligned} \tag{A.25}$$

which can be rearranged by dividing through by S_1 and collecting terms, and written,

$$\begin{aligned} & \left[1 + M\,(M+2) \right] p_1^+ + \left[1 + M\,(M-2) \right] p_1^- \\ & \quad = \left[1 + M \left(\tfrac{S_1}{S_2} M + 2 \right) \right] p_2^+ + \left[1 + M \left(\tfrac{S_1}{S_2} M - 2 \right) \right] p_2^- . \end{aligned} \tag{A.26}$$

A.3.2 Expansion

Re-evaluating the integrals in Eq. A.22, assuming uniform flow over an area equal to S_1 at the S_2 plane (zero elsewhere) and recalling that $p_3 = p_1$ as before, we have (Eq. 4.7, Davies 1988, p. 103),

$$S_1 p_1 + S_3 p_3 - S_2 p_2 = S_1 \left(\rho u_0^2 + 2 \rho_0 u_0 u \right)_2 - S_1 \left(\rho u_0^2 + 2 \rho_0 u_0 u \right)_1 . \tag{A.27}$$

Making the same substitutions for pressure, density and velocity, as before, we obtain (Eq. 4.8, Davies 1988, p. 103):

$$\begin{aligned} & (S_1 + S_3) \left(p_1^+ + p_1^- \right) - S_2 \left(p_2^+ + p_2^- \right) \\ & \quad = S_1 \left[M^2 \left(p_2^+ + p_2^- + \delta \right) + 2M \left(p_2^+ - p_2^- \right) \right] - S_1 \left[M^2 \left(p_1^+ + p_1^- \right) + 2M \left(p_1^+ - p_1^- \right) \right] \end{aligned} \tag{A.28}$$

which simplifies to,

$$\begin{aligned} & \left[S_2 + S_1 M\,(M+2) \right] p_1^+ + \left[S_2 + S_1 M\,(M-2) \right] p_1^- \\ & \quad = \left[S_2 + S_1 M\,(M+2) \right] p_2^+ + \left[S_2 + S_1 M\,(M-2) \right] p_2^- + S_1 M^2 \delta , \end{aligned} \tag{A.29}$$

and finally, dividing by S_2, we have

$$\begin{aligned} & \left[1 + \tfrac{S_1}{S_2} M\,(M+2) \right] p_1^+ + \left[1 + \tfrac{S_1}{S_2} M\,(M-2) \right] p_1^- \\ & \quad = \left[1 + \tfrac{S_1}{S_2} M\,(M+2) \right] p_2^+ + \left[1 + \tfrac{S_1}{S_2} M\,(M-2) \right] p_2^- + \tfrac{S_1}{S_2} M^2 \delta . \end{aligned} \tag{A.30}$$

Note how Eq. A.26 differs from Eq. A.30 by the additional S_1/S_2 factor in the $S_1 M p_2$ terms on the right-hand side.

A.3.3 No flow

Hence, for zero net flow ($M = 0$, $\delta = 0$), Eq. A.30 reduces to (Eq. 4.6, Davies 1988, p. 103):

$$p_1^+ + p_1^- = p_2^+ + p_2^- \,. \tag{A.31}$$

A.4 Conservation of energy

For a fixed mass of gas in a steady flow, the equation for conservation of energy q' (Bernoulli's equation) states:

$$q' = e' + \frac{p'}{\rho'} + \frac{u'^2}{2} = \text{const. (per unit mass)}, \tag{A.32}$$

where internal energy, $e' = T_{\text{stag}} s$, for entropy s; the potential (work) energy $= p'/\rho'$, for pressure p' and density ρ'; and the kinetic energy $= u'^2/2$, for velocity u'. The stagnation temperature T_{stag} is the temperature that would be obtained if the fluid were brought to rest in an adiabatic process. The enthalpy is the stored energy:

$$h' = e' + \frac{p'}{\rho'} = c_P T' \,. \tag{A.33}$$

We shall consider small perturbations about the mean values of energy, pressure, density and velocity; $e' \doteq e_0 + e$, $p' \doteq p_0 + p$, $\rho' \doteq \rho_0 + \rho$, and $u' \doteq u_0 + u$. The energy q_0, at the steady state (mean over time), is equal to the total energy q' at the perturbed state:

$$q' \;=\; e_0 + e + \frac{p_0 + p}{\rho'} + \frac{(u_0 + u)^2}{2} \,, \text{ and} \tag{A.34}$$

$$q_0 \;=\; e_0 + \frac{p_0}{\rho_0} + \frac{u_0^2}{2} \,. \tag{A.35}$$

Thus, subtracting the steady-state energy q_0 from the total q' leaves zero, by conservation of energy. Davies linearises each term in the expression by neglecting second-order terms and higher (Davies 1988), so Eq. A.35 minus Eq. A.35 becomes:

$$q' - q_0 = e + \frac{p}{\rho'} + u_0 u = 0 \,. \tag{A.36}$$

Integrating Eq. A.36 over the control volume, and recalling that $u_3 = 0$, the equation of conservation of energy for an expansion can be written,

$$\begin{aligned} \int_{S_1} (T_0 s + p/\rho' + u_0 u)_1 \, dS_1 + \int_{S_3} (T_0 s + p/\rho')_3 \, dS_3 \\ = \int_{S_2} (T_0 s + p/\rho' + u_0 u)_2 \, dS_2 \,, \end{aligned} \tag{A.37}$$

and similarly, the integral over S_3 is the only modification to the expression for a contraction. Finally, we can take care of the density ρ' in the denominators by considering the enthalpy in a compressible, adiabatic process and averaging over time and space.

A.4.1 Compressible, adiabatic, steady-flow processes

We know, from conservation of energy, that:

$$h_{\text{stag}} = h' + \frac{u'^2}{2} \tag{A.38}$$

and, for adiabatic processes, that:

$$h' = c_P T' \,, \tag{A.39}$$

where $c_P = \gamma R/(\gamma - 1)$. Substituting Eq. A.39 into Eq. A.38, gives:

$$c_P T_{\text{stag}} = c_P T' + \frac{u'^2}{2} \,. \tag{A.40}$$

Dividing by c_P, we have:

$$T_{\text{stag}} = T' \left(1 + \frac{u'^2}{2 c_P T'} \right) \,. \tag{A.41}$$

into which we can substitute for c_P, recalling Eq. A.9 in the form $c'^2 = \sqrt{\gamma R T'}$:

$$T_{\text{stag}} = T' \left(1 + \frac{1}{2} (\gamma - 1) M^2 \right) \,, \tag{A.42}$$

where $M = u'/c'$.

Now, from the adiabatic law and the equation of state (ideal gas law):

$$\frac{p_{\text{stag}}}{p'} = \left(\frac{T_{\text{stag}}}{T'} \right)^{\frac{\gamma}{\gamma - 1}} \,, \tag{A.43}$$

which, substituted into Eq. A.42, gives the expression for the local stagnation pressure, which is the pressure that would be obtained if the fluid were brought to rest in an adiabatic and reversible process:

$$p_{\text{stag}} = p' \left(1 + \frac{1}{2} (\gamma - 1) M'^2 \right)^{\frac{\gamma}{\gamma - 1}} \,, \tag{A.44}$$

and recalling Eq. A.4, it can be re-written for density:

$$\rho_{\text{stag}} = \rho' \left(1 + \frac{1}{2} (\gamma - 1) M'^2 \right)^{\frac{1}{\gamma - 1}} \,, \tag{A.45}$$

that is, the local stagnation density. Clearly, by combining Eq. A.9 with Eq. A.42, the speed of sound is also a function of the Mach number:

$$c_{\text{stag}} = c' \left(1 + \frac{1}{2} (\gamma - 1) M'^2 \right)^{\frac{1}{2}} \,, \tag{A.46}$$

which implies that:

$$M_{\text{stag}} = M' \left(1 + \frac{1}{2} (\gamma - 1) M'^2 \right)^{-\frac{1}{2}} \,. \tag{A.47}$$

This effect is small, however, and is therefore neglected.

These results, Eqs. A.42, A.44 and A.45, take account of the perturbations in density resulting from flow, but not from acoustic perturbation, since time-averaged values are used. An alternative approach, which considers the acoustic perturbations, but not the Bernoulli ones, is contained in Section A.4.5.

A.4.2 Contraction

We can derive the expression for the conservation of enthalpy at a contraction by evaluating the integrals in Eq. A.37 ($\delta = 0$, $M = M_1 = u_0/c_0$, and $M_2 = u_0 S_1/c_0 S_2 = M S_1/S_2$), and assuming the process to be isentropic (i.e., $T_0 s_1 = T_0 s_3 = T_0 s_2 = 0$). Using the standard substitutions, we get:

$$S_1 \left[\frac{p_1^+ + p_1^-}{\rho_1'} + M \frac{p_1^+ - p_1^-}{\rho_1'} \right] = S_3 \left[\frac{p_3^+ + p_3^-}{\rho_2'} \right] + S_2 \left[\frac{p_2^+ + p_2^-}{\rho_2'} + \frac{S_1}{S_2} M \frac{p_2^+ - p_2^-}{\rho_2'} \right]. \quad \text{(A.48)}$$

Recalling that for a contraction $S_1 + S_3 = S_2$, $p_3 = p_2$ and so $\rho_3' = \rho_2'$, Eq. A.48 simplifies to:

$$\frac{S_1}{\rho_1'} \left[(1+M)\, p_1^+ + (1-M)\, p_1^- \right] = \frac{S_1}{\rho_2'} \left[(1+M)\, p_2^+ + (1-M)\, p_2^- \right]. \quad \text{(A.49)}$$

However, if we consider the mean flow densities $\bar{\rho}$, averaged over the duct's cross-sectional area and over time, we can relate them using Eq. A.45 to incorporate the result for isentropic flow:

$$\frac{\bar{\rho}_2}{\bar{\rho}_1} = \left[\frac{1 - M^2 \frac{\gamma-1}{2}}{1 - \left(\frac{S_1}{S_2} M\right)^2 \left(\frac{\gamma-1}{2}\right)} \right]^{\frac{1}{\gamma-1}}, \quad \text{(A.50)}$$

then dividing Eq. A.49 through by S_2 gives us

$$\frac{\left[(1+M)\, p_1^+ + (1-M)\, p_1^- \right]}{\left[1 - \left(\frac{S_1 M}{S_2}\right)^2 \left(\frac{\gamma-1}{2}\right) \right]^{\frac{1}{\gamma-1}}} = \frac{\left[(1+M)\, p_2^+ + (1-M)\, p_2^- \right]}{\left[1 - M^2 \left(\frac{\gamma-1}{2}\right) \right]^{\frac{1}{\gamma-1}}}. \quad \text{(A.51)}$$

A.4.3 Expansion

If we now consider that the left-hand side of Eq. A.37 is isentropic (zero net entropy flux), i.e., $T_0 s_1 = T_0 s_3 = 0$, and we use the term derived in Eq. 2.4 (Davies 1988, p. 94) for the right-hand side, $T_0 s_2 = -\delta/\rho_0(\gamma - 1)$, we can re-evaluate the integrals to obtain the expression at an expansion:

$$\begin{aligned} S_1 \left[\tfrac{p_1^+ + p_1^-}{\rho_1'} + M \tfrac{(p_1^+ - p_1^-)}{\rho_1'} \right] + S_3 \left[\tfrac{p_3^+ + p_3^-}{\rho_3'} \right] \\ = S_2 \left[\tfrac{-\delta}{\rho_2'(\gamma-1)} + \tfrac{p_2^+ + p_2^-}{\rho_2'} + \tfrac{S_1}{S_2} M \tfrac{(p_2^+ - p_2^-)}{\rho_2'} \right]. \end{aligned} \quad \text{(A.52)}$$

We recall that $p_3 = p_1$, hence $\rho_3' = \rho_1'$, and that $S_1 + S_3 = S_2$. So, dividing through by S_2, as before, and rearranging, this gives

$$\frac{\left[\left(1 + \frac{S_1}{S_2} M\right) p_1^+ + \left(1 - \frac{S_1}{S_2} M\right) p_1^- \right]}{\rho_1'} = \frac{\left[\left(1 + \frac{S_1}{S_2} M\right) p_2^+ + \left(1 - \frac{S_1}{S_2} M\right) p_2^- - \frac{\delta}{\gamma-1} \right]}{\rho_2'}, \quad \text{(A.53)}$$

which is identical to Eq. A.49 for $\delta = 0$, but differs from Eq. 4.5 (Davies 1988, p. 102), replacing M with MS_1/S_2. Incorporating the temporally- and spatially-averaged flow densities from Eq. A.50 again, the expression becomes,

$$\frac{\left[\left(1 + \frac{S_1}{S_2} M\right) p_1^+ + \left(1 - \frac{S_1}{S_2} M\right) p_1^- \right]}{\left[1 - \left(\frac{S_1}{S_2} M\right)^2 \left(\frac{\gamma-1}{2}\right) \right]^{\frac{1}{\gamma-1}}} = \frac{\left[\left(1 + \frac{S_1}{S_2} M\right) p_2^+ + \left(1 - \frac{S_1}{S_2} M\right) p_2^- - \frac{\delta}{\gamma-1} \right]}{\left[1 - M^2 \left(\frac{\gamma-1}{2}\right) \right]^{\frac{1}{\gamma-1}}}. \quad \text{(A.54)}$$

This result accounts for the effects of mixing, which produce the fully-developed flow further downstream.

A.4.4 No flow

Hence, for zero net flow ($M = 0$, $\delta = 0$), Eq. A.54 also reduces to (Eq. 4.6, Davies 1988, p. 103):

$$p_1^+ + p_1^- = p_2^+ + p_2^- \,. \tag{A.55}$$

which is identical to Eq. A.31, derived earlier from momentum.

A.4.5 Linearisation of pressure upon density

In this section, an alternative linearisation of the pressure upon density to that of Davies (1988) is formulated. In contrast, it includes an additional term for the simultaneous interaction of the acoustic wave with the local pressure and density, but it does not account for the effect of flow on the adiabatic process, as in Section A.4.1.

Treating the effects of the density perturbation as a binomial expansion, the expression for the ratio of the perturbed pressure to the perturbed density in Eq. A.35 becomes:

$$\begin{aligned}
\tfrac{p_0+p}{\rho_0+\rho} &= \tfrac{p_0}{\rho_0}\left(1+\tfrac{\rho}{\rho_0}\right)^{-1} + \tfrac{p}{\rho_0}\left(1+\tfrac{\rho}{\rho_0}\right)^{-1} \\
&= \tfrac{1}{\rho_0}\left[p_0\left(1-\tfrac{\rho}{\rho_0}+\left(\tfrac{\rho}{\rho_0}\right)^2-\left(\tfrac{\rho}{\rho_0}\right)^3+\ldots\right)+p\left(1-\tfrac{\rho}{\rho_0}+\left(\tfrac{\rho}{\rho_0}\right)^2-\left(\tfrac{\rho}{\rho_0}\right)^3+\ldots\right)\right]
\end{aligned}$$

To linearise, we neglect second order terms and higher, which leaves:

$$\tfrac{p_0+p}{\rho_0+\rho} \approx \tfrac{1}{\rho_0}\left[p_0\left(1-\tfrac{\rho}{\rho_0}\right)+p\right].$$

Therefore, by subtracting the steady state and recalling $c_0^2 = \gamma p_0/\rho_0$, the change in potential energy (PE) can be expressed as:

$$\begin{aligned}
\Delta(\mathrm{PE}) &= \frac{p_0+p}{\rho_0+\rho} - \frac{p_0}{\rho_0} \\
&= \frac{1}{\rho_0}\left(p - \frac{c_0^2}{\gamma}\rho\right), .
\end{aligned} \tag{A.56}$$

Now, making the usual substitutions for density in regions 1 and 3, $\rho = p/c_0^2$,

$$\begin{aligned}
\Delta(\mathrm{PE})_1 &= \frac{p_1}{\rho_0}\left(1-\frac{1}{\gamma}\right), \text{ and} \\
\Delta(\mathrm{PE})_3 &= \frac{p_3}{\rho_0}\left(1-\frac{1}{\gamma}\right);
\end{aligned}$$

and in region 2, $\rho_2 = (p_2+\delta)/c_0^2$,

$$\Delta(\mathrm{PE})_2 = \frac{p_2}{\rho_0}\left(1-\frac{1}{\gamma}\right) - \frac{\delta}{\rho_0\gamma}\,.$$

Hence, the reformulation of the equation of conservation of energy is:

$$\begin{aligned}\int_{S_1}\left(T_0 s_1 + \tfrac{p_1}{\rho_0}\left(1-\tfrac{1}{\gamma}\right) + u_0 u_1\right) dS_1 + \int_{S_3}\left(T_0 s_3 + \tfrac{p_3}{\rho_0}\left(1-\tfrac{1}{\gamma}\right)\right) dS_3 \\ = \int_{S_2}\left(T_0 s_2 + \tfrac{p_2}{\rho_0}\left(1-\tfrac{1}{\gamma}\right) - \tfrac{\delta}{\rho_0\gamma} + u_0 u_2\right) dS_2 \,.\end{aligned} \tag{A.57}$$

which contains extra terms when compared with the earlier expression, Eq. A.37, that ignored the acoustic density fluctuation.

A.5 Side branch

When modelling geometries more complex that a single duct, the plane-wave formulation can be extended to calulate the transfer at the interface of the main tract with any side branches. We first demonstrate how the equations for no flow can be derived using the classical electrical analogue, then we expand these to allow for flow by applying the conservation laws to a contraction.

A.5.1 No flow

If we choose to ignore the effects of net flow, we can describe the pressure transfer at an abrupt area change with a side branch (or sinus) by reference to the electrical analogy. Thus, acoustic admittances at a junction sum to zero so, for the contraction geometry in Figure A.3a. We can write

$$\frac{S_2}{z_2} = \frac{S_1}{z_1} + \frac{S_3}{z_3}\,, \tag{A.58}$$

where z_i is the acoustic impedance in region i (defined in Eq. A.11). Also note that the areas

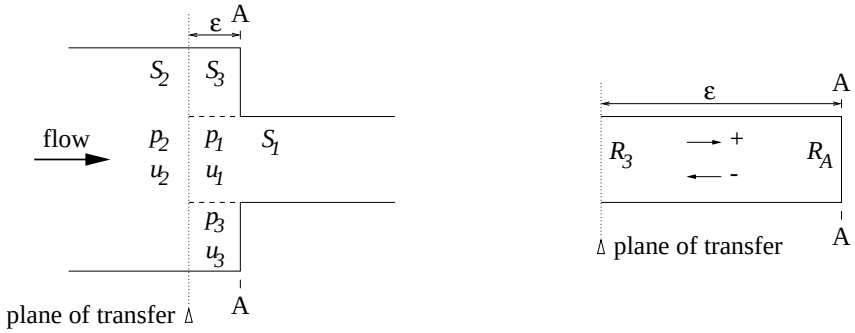


Figure A.3: Tube interfaces indicating the pressures, velocities and areas used in the calculation of reflection coefficients for (a) a contraction, and (b) a tube closed at one end.

either side of the place of transfer are equal,

$$S_2 = S_1 + S_3\,. \tag{A.59}$$

The reflection coefficient R_A at the closed end of the side branch (at A) is the ratio of the reflected pressure to the incident pressure:

$$R_A = \frac{p_A^-}{p_A^+}, \tag{A.60}$$

where p_A^+ and p_A^- are related to p_3^+ and p_3^- by the propagation time over the branch length x_3:

$$\begin{aligned} p_3^+ &= p_A^+ \exp(+jk\epsilon) \\ p_3^- &= p_A^- \exp(-jk\epsilon) \end{aligned} \tag{A.61}$$

In practice, the reflection coefficient, $R_A \approx 1$, depends on the absorption of the end wall. Hence, the reflection coefficient at the transfer plane is

$$\begin{aligned} R_3 &= \frac{p_3^-}{p_3^+} \\ &= R_A \exp(-j2k\epsilon). \end{aligned} \tag{A.62}$$

Re-writing R_3 in terms of the acoustic impedance, using Eq. A.11, and replacing the partial pressures with $p_3 = p_3^+ + p_3^-$ and $u_3 = (p_3^+ - p_3^-)/\rho_0 c_0$, we have

$$\begin{aligned} z_3 &= \frac{p_3}{u_3} \\ &= \rho_0 c_0 \frac{p_3^+ + p_3^-}{p_3^+ - p_3^-} \\ &= \rho_0 c_0 \frac{1 + R_3}{1 - R_3} \\ &= \rho_0 c_0 \frac{1 + R_A \exp(-j2k\epsilon)}{1 - R_A \exp(-j2k\epsilon)}. \end{aligned} \tag{A.63}$$

Now, from previously calculated pressure, the impedance z_1 in the narrow tube (region 1) is

$$z_1 = \rho_0 c_0 \frac{1 + p_1^-/p_1^+}{1 - p_1^-/p_1^+}. \tag{A.64}$$

Recalling Eq. A.58, we get a final expression for the acoustic impedance of region 2, which relates the partial pressures p_2^+ and p_2^-:

$$z_2 = \frac{\rho_0 c_0}{\frac{1+p_1^-/p_1^+}{1-p_1^-/p_1^+} + \frac{1+R_A \exp(-j2k\epsilon)}{1-R_A \exp(-j2k\epsilon)}}. \tag{A.65}$$

To find the transfer, let us equate pressures according to conservation of momentum:

$$p_1^+ + p_1^- = p_2^+ + p_2^- \tag{A.66}$$

$$= p_3^+ + p_3^- \tag{A.67}$$

and equate volume velocity by continuity of mass,

$$S_2 u_2 = S_1 u_1 + S_3 u_3 \tag{A.68}$$

Also, for plane waves,

$$u_1 = \frac{p_1^+ - p_1^-}{\rho_0 c_0}$$
$$u_2 = \frac{p_2^+ - p_2^-}{\rho_0 c_0}$$
$$u_3 = \frac{p_3^+ - p_3^-}{\rho_0 c_0} \tag{A.69}$$

Now, substituting for pressure and multiplying by $\rho_0 c_0 / S_2$, Eq. A.68 becomes

$$\left(p_2^+ - p_2^-\right) = \frac{S_1}{S_2}\left(p_1^+ - p_1^-\right) + \frac{S_3}{S_2}\left(p_3^+ - p_3^-\right) , \tag{A.70}$$

but,

$$p_3^+ - p_3^- = \rho_0 c_0 \frac{p_3^+ + p_3^-}{z_3} . \tag{A.71}$$

Recalling Eq. A.67 and substituting into Eq. A.70 thus gives

$$\left(p_2^+ - p_2^-\right) = \frac{S_1}{S_2}\left(p_1^+ - p_1^-\right) + \frac{S_3}{S_2}\frac{\rho_0 c_0}{z_3}\left(p_1^+ + p_1^-\right) . \tag{A.72}$$

Adding Eqs. A.66 and A.72 and rearranging produces the result

$$p_2^+ = \frac{1}{2}\left[(p_1^+ + p_1^-) + \frac{S_1}{S_2}(p_1^+ - p_1^-) + \frac{S_3}{S_2}\frac{\rho_0 c_0}{z_3}(p_1^+ + p_1^-)\right] , \tag{A.73}$$

and subtracting Eq. A.66 from Eq. A.72 leaves

$$p_2^- = \frac{1}{2}\left[(p_1^+ + p_1^-) - \frac{S_1}{S_2}(p_1^+ - p_1^-) - \frac{S_3}{S_2}\frac{\rho_0 c_0}{z_3}(p_1^+ + p_1^-)\right] . \tag{A.74}$$

Rearranging to separate the contributions from the positive and negative travelling waves, and substituting $\Sigma = (S_1/S_2)$ and $\Omega = (\rho_0 c_0 S_3 / z_3 S_2)$:

$$\begin{aligned} p_2^+ &= \frac{1}{2}\left[p_1^+ (1 + \Sigma + \Omega) + p_1^- (1 - \Sigma + \Omega)\right] \\ p_2^- &= \frac{1}{2}\left[p_1^+ (1 - \Sigma - \Omega) + p_1^- (1 + \Sigma - \Omega)\right] \end{aligned} \tag{A.75}$$

A.5.2 Steady flow

In the case of a net flow, assuming Mach number $M = M_2 = M_1$, elsewhere $M_3 = 0$, the linearised equation for continuity of mass (Eq. A.72) becomes,

$$S_2\left[(1+M)p_2^+ - (1-M)p_2^-\right] = S_3(p_3^+ - p_3^-) + S_1\left[(1+M)p_1^+ - (1-M)p_1^-\right] ,$$

or rearranging,

$$(1+M)p_2^+ - (1-M)p_2^- = \frac{S_3}{S_2\zeta_3}(p_3^+ + p_3^-) + \frac{S_1}{S_2}\left[(1+M)p_1^+ - (1-M)p_1^-\right] . \tag{A.76}$$

Considering the linearised formulae for the conservation of energy, Eqs. A.66 and A.67 become

$$\begin{aligned} (1+M)p_2^+ + (1-M)p_2^- &= p_3^+ + p_3^- \\ &= (1+M)p_1^+ + (1-M)p_1^- . \end{aligned} \tag{A.77}$$

Now, making the substitutions $\Sigma = (S_1/S_2)$ and $\Omega = (\rho_0 c_0 S_3/z_3 S_2)$, as before, and adding Eq. A.77 to Eq. A.76 yields

$$\begin{aligned} p_2^+ &= \tfrac{1}{2(1+M)}\left[(1+M)p_1^+ + (1-M)p_1^- + \Sigma\left((1+M)p_1^+ - (1-M)p_1^-\right) + \Omega(p_1^+ + p_1^-)\right] \\ &= \tfrac{1}{2(1+M)}\left[p_1^+\left(1+M+\Sigma(1+M)+\Omega\right) + p_1^-\left(1-M-\Sigma(1-M)+\Omega\right)\right]. \end{aligned} \tag{A.78}$$

Subtracting Eq. A.76 from Eq. A.77, we obtain

$$\begin{aligned} p_2^- &= \tfrac{1}{2(1-M)}\left[(1+M)p_1^+ + (1-M)p_1^- - \Sigma\left((1+M)p_1^+ - (1-M)p_1^-\right) - \Omega(p_1^+ + p_1^-)\right] \\ &= \tfrac{1}{2(1-M)}\left[p_1^+\left(1+M-\Sigma(1+M)-\Omega\right) + p_1^-\left(1-M+\Sigma(1-M)-\Omega\right)\right]. \end{aligned} \tag{A.79}$$

Similar expressions can be obtained for an expansion, in the same way.

A.6 Note on radiation impedance

Rather like the end-correction factors at abrupt area changes, the termination at the open end of the vocal tract, treated as a piston in an infinite baffle, can also be adjusted by extending the length of the tract in the plane-wave model. By fitting a curve to the calculated response and experimental results, the following approximation has been derived (Davies et al. 1980; Davies 1988; Munjal 1987):

$$\epsilon = \left(b_0 r - b_1 kr - b_2 (kr)^2\right)\left(1 - M^2\right), \tag{A.80}$$

where $0 < M < 0.4$ is the Mach number in the duct, whose hydraulic radius is r, $k = 2\pi f/c_0$ is the wavenumber, and the constants, b_0, b_1 and b_2, take the values

$$b_0,\ b_1,\ b_2 = \begin{cases} 0.6133,\ 0,\ 0.1168 & \text{for } kr < 0.5; \\ 0.6393,\ 0.1104,\ 0 & \text{for } 0.5 < kr < 2. \end{cases} \tag{A.81}$$

This expression can be used directly to calculate the effective radiation impedance at the lips, and hence the reflection coefficient. Note that Eq. A.81 is similar to that for an end correction, which is given in Section 2.3.6. The end-correction equation is, however, different from this expression for the open end (Davies 1988, Eq. 3.10, p. 99), which depends on flow and on frequency.

A.7 Intermediate source in a simple tube

This section provides an illustrative example of the transfer function calculation for an ideal pressure source in a simple rigid tube, closed at one end, the 'glottis'. It gives a proof that the transfer function from source to the aperture, or 'lips', is equal to the ratio of two other

transfer functions: that from the glottis to the lips, divided by that from the glottis to the source. This evidence is relevant to the explanation and development of Section 2.4.2.

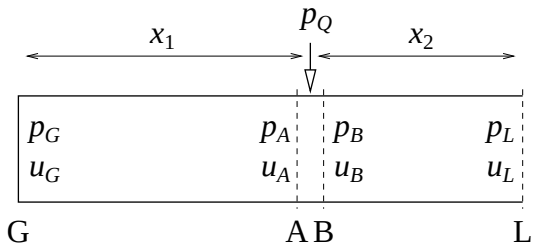


Figure A.4: Intermediate pressure source at Q in a simple tube that is closed at the left-hand end G and open at the other end L.

According to the locations in Figure A.4, we can write the sound pressures as the sum of the positive- and negative-travelling plane-wave components:

$$\begin{aligned} p_G &= p_G^+ + p_G^- \\ p_A &= p_A^+ + p_A^- \\ p_B &= p_B^+ + p_B^- \\ p_L &= p_L^+ + p_L^-; \end{aligned} \tag{A.82}$$

the velocities are proportional to the differences:

$$\begin{aligned} u_G &= \left(p_G^+ - p_G^-\right)/\rho_0 c_0 \\ u_A &= \left(p_A^+ - p_A^-\right)/\rho_0 c_0 \\ u_B &= \left(p_B^+ - p_B^-\right)/\rho_0 c_0 \\ u_L &= \left(p_L^+ - p_L^-\right)/\rho_0 c_0 \,. \end{aligned} \tag{A.83}$$

Calculating the transfer from the glottis to A, which is just to the left of the source, gives

$$\begin{aligned} p_A^+ &= p_G^+ \exp -jkx_1 \\ p_A^- &= p_G^- \exp +jkx_1 \,, \end{aligned} \tag{A.84}$$

then, traversing the pressure source p_Q to B, just to the right of Q, gives

$$\begin{aligned} p_B^+ &= \frac{p_Q}{2} + p_A^+ \\ &= \frac{p_Q}{2} + p_G^+ \exp -jkx_1 \,; \end{aligned} \tag{A.85}$$

$$\begin{aligned} p_B^- &= -\frac{p_Q}{2} + p_A^- \\ &= -\frac{p_Q}{2} + p_G^- \exp +jkx_1 \,. \end{aligned} \tag{A.86}$$

At the lips, the relationship between pressure and velocity is defined by the radiation impedance, thus:

$$z_L = \frac{p_L}{u_L} = \rho_0 c_0 \frac{(1+R_L)}{(1-R_L)}\,, \tag{A.87}$$

where the reflection coefficient $R_L = p_L^-/p_L^+$, and the partial pressures are

$$\begin{aligned} p_L^+ &= p_B^+ \exp -jkx_2 \\ &= \frac{p_Q}{2} \exp -jkx_2 \; + \; p_G^+ \exp -jk\,(x_1 + x_2) \; ; \end{aligned} \tag{A.88}$$

$$\begin{aligned} p_L^- &= p_B^- \exp +jkx_2 \\ &= -\frac{p_Q}{2} \exp +jkx_2 \; + \; p_G^- \exp +jk\,(x_1 + x_2) \; . \end{aligned} \tag{A.89}$$

Let us derive the simple transfer functions in which we are interested. First, the glottis-source TF, which is from a notional velocity at G to the pressure at Q, is

$$H_{GQ}^P = \left(\frac{p_Q}{u_G}\right)_{u_Q = 0} \tag{A.90}$$

and we note that $p_A^+ = p_A^- = p_Q/2$; hence,

$$\begin{aligned} H_{GQ}^P &= \frac{p_Q \rho_0 c_0}{\frac{p_Q}{2}\left(\exp +jkx_1 \; - \; \exp -jkx_1\right)} \\ &= \frac{2\rho_0 c_0}{\exp +jkx_1 \; - \; \exp -jkx_1} \, . \end{aligned} \tag{A.91}$$

Second, the glottis-lips TF, which is from a notional velocity at G to the velocity at L, is

$$\begin{aligned} H_{GL}^V &= \left(\frac{u_L}{u_G}\right)_{u_L = (1-R_L)p_L^+} \\ &= \frac{\left(p_L^+ - p_L^-\right)}{\left(p_L^+ \exp +jk\,(x_1 + x_2) \; - \; p_L^- \exp -jk\,(x_1 + x_2)\right)} \frac{\rho_0 c_0}{\rho_0 c_0} \\ &= \frac{(1 - R_L)}{\exp +jk\,(x_1 + x_2) \; - \; R_L \exp -jk\,(x_1 + x_2)} \, . \end{aligned} \tag{A.92}$$

Now since it has a closed end, let us assume that there is no acoustic velocity at the glottis, i.e., $p_G^+ - p_G^- = 0$:

$$\begin{aligned} \Rightarrow \quad & p_L^+ \exp +jk\,(x_1 + x_2) \; - \; \tfrac{p_Q}{2} \exp -jkx_2 \exp +jk\,(x_1 + x_2) \\ & \qquad = p_L^- \exp -jk\,(x_1 + x_2) \; + \; \tfrac{p_Q}{2} \exp +jkx_2 \exp -jk\,(x_1 + x_2) \; , \end{aligned} \tag{A.93}$$

and recalling the reflection coefficient,

$$\begin{aligned} \Rightarrow \quad & p_L^+ \left[\exp +jk\,(x_1 + x_2) \; - \; R_L \exp -jk\,(x_1 + x_2)\right] \\ & \qquad = \tfrac{p_Q}{2} \left[\exp -jkx_1 - \exp +jkx_1\right] \, . \end{aligned} \tag{A.94}$$

So, the transfer function from the source to lips is

$$\begin{aligned} H_{QL}^P(\omega) &= \tfrac{u_L}{p_Q} \\ &= \tfrac{p_L^+(1-R_L)}{p_Q \rho_0 c_0} \\ &= \tfrac{\exp -jkx_1 - \exp +jkx_1}{2\rho_0 c_0} \quad \tfrac{1-R_L}{\exp +jk(x_1+x_2) - R_L \exp -jk(x_1+x_2)} \\ &= H_{QG}^P(\omega) \qquad\qquad H_{GL}^V(\omega) \, . \end{aligned} \tag{A.95}$$

QED, by comparison with Eqs. A.91 and A.92.

Appendix B

VOAC pseudo-code transcription

B.1 Testing

This appendix contains an account of early tests on the translated VOAC program, and describes the current input file format with an illustrative example. Afterward, a pseudo-code transcription of the current version of VOAC (v5.1.3) is given. Figure B.1 below depicts the structure of the program, whose subroutines correspond to the later sections of this appendix.

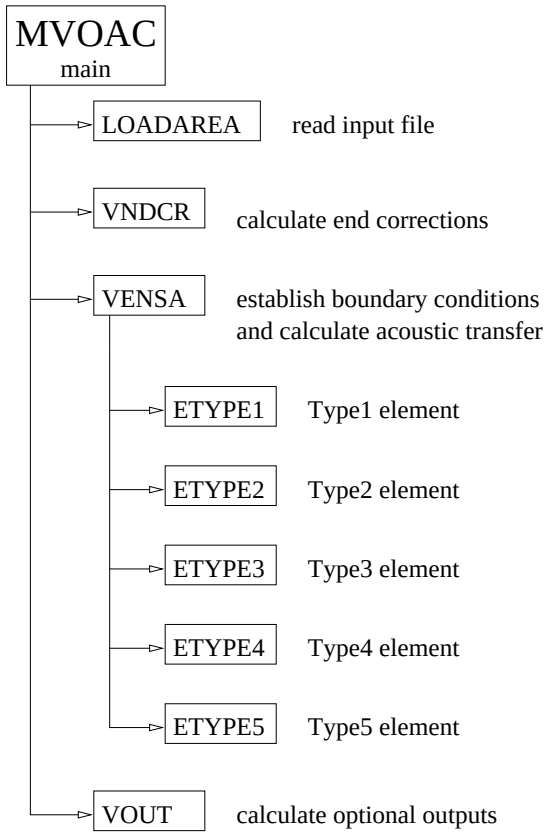


Figure B.1: Program structure of the current version of VOAC: v5.1.0 (for Matlab 4.2).

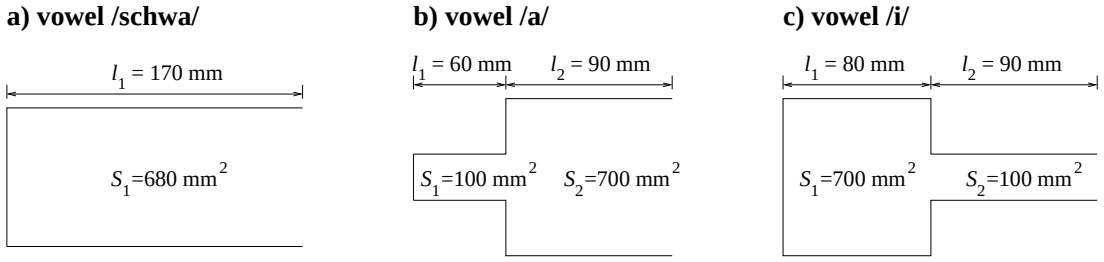


Figure B.2: Tube representations of vowel geometries: (a) single-tube test /ə/, and two-tube tests, (b) /ɑ/ and (c) /i/.

B.1.1 Preliminary and system tests

The first set of test input files were designed to exercise each major section of code within each module as far as possible, with the exception of the expansion (Type 4), which was only tested without any side branches (i.e., $x_2 = x_4 = 0$). Initially, the Matlab version was tested for consistency against its Fortran ancestor. Simple, single-element input files were created for each element type to represent a tube of length $l = 17\,\mathrm{cm}$, closed at one end. In the tests, the two versions of the program were brought into alignment, such that their results were identical (to within the limit of numerical accuracy). However, it was discovered that the area change within Type 2 elements could not be set exactly equal to zero (i.e., $\alpha \neq 0$). Once all program components had successfully completed the test procedure, the input file representing the Fant /ɨ/ was applied. As before, the results of the two programs were compared and, for this more complicated example, found to be consistent. The tests were repeated successfully using non-zero values for wall compliance. Similarly, in the second set of tests, small perturbations were made to a specified geometry to test as many different paths in the code as possible. The test results were based on the computed formant frequencies, which facilitated the detection of bugs in the source code.

B.1.2 Formant frequencies

In order to test VOAC more thoroughly and to relate the results back to analytical calculations of the acoustical response, three simple tube geometries were employed, crudely representing the vowels /ə/, /ɑ/ and /i/, as drawn in Figure B.2. VOAC offers a choice of quantities to plot as output. For these tests, the volume-velocity transfer function $H^V(f)$ was used, which was computed over the range 20–5000 Hz at 10 Hz intervals. The peaks in the magnitude response of the transfer function were picked as estimates of the resonance frequencies, or formants.

For the single-tube test representing /ə/, hand-calculation of the resonance frequencies f_n was done from the expression for plane, standing waves in a rigid, lossless (ideal) tube, closed

at one end:

$$f_n = \frac{c_0}{4l}(2n+1)\,, \tag{B.1}$$

where c_0 is the speed of sound, l is the tube length and n a natural number. The two-tube tests required the additional assumption that the abrupt area changes were large (i.e., $S_1 \ll S_2$ for /ɑ/ and $S_1 \gg S_2$ for /i/). Hence, a rough estimate of the resonance frequencies can be made by decoupling the two tubes and using expressions for standing waves. In the case of /ɑ/, both sections of the tube can be considered closed-open, whereas for /i/ the first is effectively closed-closed (left) and the second open-open (right). The formula for a section of tube open at both ends, or closed at both ends is:

$$f_n = \frac{c_0 n}{2l}\,. \tag{B.2}$$

The Helmholtz resonance was included for /i/, which has a volume of air $S_1 l_1$ enclosed by a narrow neck of length l_2 and area S_2:

$$f_H = \frac{c_0}{2\pi}\sqrt{\frac{S_2}{S_1 l_1 l_2}}\,. \tag{B.3}$$

The first few values of the formant frequencies (in Hz) were calculated as:

/ə/: 500, 1500, 2500, 3500, 4500, 5500, ...
/ɑ/: 750, 1250, 2700, 3280, 4860, 5470, ...
/i/: 270, 1940, 2920, 3890, 5830, 5830, ...

By comparing these values with the test results, the resonance frequencies give an instant and plain indication of whether the program is working correctly or not.

B.1.3 Secondary tests

Nominal values were computed by VOAC, incorporating the effects of end corrections and radiation impedance, as follows (Hz, the percentage difference from the hand-calculated values is shown in brackets):

/ə/: 500 (+0.0 %), 1510 (+0.7 %), 2520 (+0.8 %), 3530 (+0.9 %), 4560 (+1.3 %)
/ɑ/: 760 (+1.3 %), 1220 (−2.4 %), 2750 (+1.9 %), 3230 (−1.5 %), 4760 (−2.1 %)
/i/: 270 (+0.0 %), 1960 (+1.0 %), 2790 (−4.5 %), 4080 (+4.9 %)

These values were computed using a Type 5 element to represent /ə/, a combination of Types 1 and 5 for /ɑ/, and Type 4 for /i/.

The first time these tests were carried out, the results varied widely (> 20 %) and it was clear that the discrepancies had been caused by a bug in the program. Indeed, in some cases there were no resonance peaks in the specified frequency range, and sometimes the program even crashed, giving no results. Therefore, taking into account the small variations in geometry

Type	1	1-5	2	2-5	3	3-5	4	4-5	5
/ə/	∼	×	√	−	×	−	√	−	√
/ɑ/	√	∼	−	√	−	×	(×)	√	−
/i/	×	−	−	√	−	×	√	−	−

Table B.1: Summary of two-tube test results.

that were introduced, and the size of the errors between the hand-calculated values and the nominal (VOAC) ones, a configuration was considered to have passed the test if its formants were within ±5 % of these nominal values.

B.1.4 Summary of test results

An exhaustive list of tests would be too lengthy (and tedious) to insert here, particularly as they owe much of their meaning to a close reading of the source code. Nevertheless, the RAMP (Type 2) tests require a special mention. The test conditions for the CONE (Type 3) were very similar, but since they all failed, their details are of less interest.

For the vowel /ə/, a variety of representations was used with the inclusion of a minor area change (∼ 1 %, about halfway along the tube) to engage different parts of the program code. For the two-tube RAMP tests, the Type 2 element was centred on the interface, with the remaining parts as concatenated PIPEs (Type 5). Thus, the abrupt area change gained a finite length, which was tested over a range (1–4 cm), the steepest configuration giving the most accurate results, as might be expected.

During testing, a number of mistakes was identified in the code and corrected. Table B.1 summarises the results of all the two-tube tests after implementation of the fixes (described in the following section), where "√" denotes pass, "×" fail, "−" no test and "∼" inconsistent behaviour. Where the geometry could not be represented by a single element, a PIPE (Type 5) was adjoined to make a two-element combination (e.g., Type 2 → Type 2-5).

The tests showed that Types 2, 4 and 5 provide valid results for expanding, contracting and constant-area geometries (the bracketed cross (×) indicates an illegal configuration for Type 4). The spherical wave-front calculation for conical sections (Type 3) was consistently incorrect, but the Type 1 element gives valid results for certain configurations. The source of these intermittent failures is to be investigated. Meanwhile, orifice geometries can be alternatively represented by the validated element types (viz. Types 2, 4 & 5).

B.1.5 Modifications

All the discrepancies, or bugs, that were found in the Matlab program code and the Fortran version were due to translation errors, except for two typographical errors in the Fortran version,

which were corrected satisfactorily (see Jackson 1997, Section 2.3, for details). Inspection of the code, following the two-tube tests, brought to light another couple of bugs and some cases of numerical instability, which occurred when the results of certain mathematical operations went to `NaN` (not a number) or `inf` (infinity).

The bugs, both in the Type 1 code, occurred for incomplete elements (i.e., using less than three sub-elements). The first bug was related to the resolution of a set of nested conditions, which resulted in the wrong sub-element's transfer pressure *qpI(f)* being used to calculate the element transfer pressure *pI(f)* (lines 195–205). The second was a transcription error in the denominator of the side-branch reflection coefficient *m(f)*, which was numerically unstable for cases of constant hydraulic radius, $r_1 = r_2$ and $r_3 = r_4$ (l. 18, l. 152). An ill-defined *m(f)* went on to destabilise later calculations (l. 43, l. 44). Further instabilities were discovered in the Type 2 code for cases of zero area change (l. 25, l. 33, l. 35). The bugs, as far as they have been understood, have been corrected, but no safeguard has been put in place against potential numerical instabilities.

B.1.6 Summary: function and dysfunction

The implication of the test results is that VOAC's current functionality is limited to Types 1, 2, 4 and 5, which merely means that the CONE element is excluded. Thus, there is little restriction of the choice of elements and indeed all area profiles can be approximated using a suitable selection of the available types.

B.2 Data format

The information required for acoustic predictions is the axial distribution of vocal-tract area S (area function) and cross-sectional shape, which is represented by the hydraulic radius r_H. Within the element definitions, it is safer always to repeat values of S and r_H when the corresponding length $x = 0$, to ensure the correct program execution. A point to note concerning the Type 4 element is that S_2 and S_4 are the total side-branch area, while r_2 and r_4 are the average hydraulic radii, when combining more than one side branch. The lowest frequency `fMin` must not be zero.

B.2.1 File contents

The variables contained in a typical data input file are:

```
>> clear; fanti; whos

  Name          Size         Bytes   Class
```

```
  Comment         1x79          158   char array
  c0              1x1             8   double array
  elArea          7x5           280   double array
  elHRad          7x5           280   double array
  elLength        7x4           224   double array
  elType          7x1            56   double array
  fMax            1x1             8   double array
  fMin            1x1             8   double array
  fStr            1x5            10   char array
  nE              1x1             8   double array
  nF              1x1             8   double array
  p0              1x1             8   double array
  qM              1x1             8   double array
  qV              1x1             8   double array
  smm             1x1             8   double array

Grand total is 198 elements using 1080 bytes
```

The `nE`-point vector of element types is `elType`, and their corresponding lengths, areas and hydraulic radii are `elLength`, `elArea` and `elHRad`, respectively. The frequency range and resolution are defined by `fMin`, `fMax` and the number of frequencies `nF`.

The fluid properties are described by the speed of sound `c0`, the ambient pressure `p0`, the mass flow-rate `qM` and volume flow-rate `qV`. The wall mass per unit area is `smm`. `Comment` contains text describing the file contents and history, and `fStr` is a text label.

B.2.2 Example file: Fant /ɨ/

The contents of a typical data input file are given below. In this case, it contains a seven-element approximation of the area function for /ɨ/, published by Fant (1960).

```
fStr='fanti';
nE=7;
elType=[1 3 2 2 1 3 1].';
elLength=[ ...
  +9.00e-03   +5.00e-03   +1.90e-02   +0.00e-01;
  +0.00e-01   +2.00e-02   +0.00e-01   +0.00e-01;
  +0.00e-01   +2.00e-02   +0.00e-01   +0.00e-01;
  +0.00e-01   +1.80e-02   +0.00e-01   +0.00e-01;
  +2.00e-03   +6.00e-03   +0.00e-01   +0.00e-01;
  +0.00e-01   +3.90e-02   +0.00e-01   +0.00e-01;
  +3.10e-02   +2.10e-02   +0.00e-01   +0.00e-01];
```

```
elLength=reshape(elLength,4,7).';

elArea=[ ...
  +6.80e-04   +3.10e-04   +8.00e-04   +6.80e-04   +0.00e-01;
  +6.80e-04   +2.00e-04   +0.00e-01   +0.00e-01   +0.00e-01;
  +2.00e-04   +1.00e-04   +0.00e-01   +0.00e-01   +0.00e-01;
  +1.00e-04   +2.50e-04   +0.00e-01   +0.00e-01   +0.00e-01;
  +2.50e-04   +1.55e-04   +3.40e-04   +3.40e-04   +0.00e-01;
  +3.40e-04   +1.14e-03   +0.00e-01   +0.00e-01   +0.00e-01;
  +1.14e-03   +2.90e-04   +2.90e-04   +2.90e-04   +0.00e-01];
elArea=reshape(elArea,5,7).';

elHRad=[ ...
  +1.30e-02    +5.00e-03    +1.50e-02    +1.20e-02    +0.00e-01;
  +1.20e-02    +6.50e-03    +0.00e-01    +0.00e-01    +0.00e-01;
  +6.50e-03    +3.20e-03    +0.00e-01    +0.00e-01    +0.00e-01;
  +3.20e-03    +7.00e-03    +0.00e-01    +0.00e-01    +0.00e-01;
  +7.00e-03    +4.00e-03    +9.00e-03    +9.00e-03    +0.00e-01;
  +9.00e-03    +1.90e-02    +0.00e-01    +0.00e-01    +0.00e-01;
  +1.90e-02    +9.50e-03    +9.50e-03    +9.50e-03    +0.00e-01];
elHRad=reshape(elHRad,5,7).';
fMin=20;
fMax=5000;
nF=499;
c0=359;
qM=1.317000e-3/6;% kg/s
p0=760;
qV=1.2/6;% litres/s
Comment=newcmnt('','SLASHI');

% Program parameters
smm=0;
```

B.3 Pseudocode

main

item	name	value	description
α	`at`	3.15×10^{-5}	unspecified constant
γ_{wet}	`gamma`	1.396	ratio of specific heats, 100% humidity
γ_{dry}	`gamma`	1.400	ratio of specific heats, 0%
c_{wet}	`c0`	359	speed of sound (ms^{-1}), 100% humidity
c_{dry}	`c0`	353	speed of sound (ms^{-1}), 0% humidity
ρ	`rho`	.	density of air (kg m^{-3})
$\rho c\|_{wet}$	`rhoc`	394	characteristic impedance (kg $\mathrm{m}^{-2}\mathrm{s}^{-1}$), 100% humidity
$\rho c\|_{dry}$	`rhoc`	402	characteristic impedance (kg $\mathrm{m}^{-2}\mathrm{s}^{-1}$), 0% humidity
T_0	.	310	ambient temperature (°K)
	`imk`	0	a constant,
a	`a`	1.5	empirical constant (end-correction),
κ	`kappa`	0.63	empirical constant (end-correction),
Δ_0	`tolZero`	1×10^{-9}	tolerance on zero,
Δ_r	`tolHRad`	0.002	tolerance on hydraulic radius (m),
Δ_θ	`tolAlpha`	0.1	tolerance on cone angle, θ.

item	name	description
N	`nE`,	number of elements,
i	`iE`,	element index,
j	`iL`,	atom (or sub-element) index,
E_i	`elType`,	element type ($E \in \{1, \cdots, 5\}$),
$l_{i,j}$	`elLength`,	elemental length (m),
$A_{i,j}$	`elArea`,	elemental area (m^2),
$r_{i,j}$	`elHRad`,	elemental hydraulic radius (m),
f_{min}	`fMin`,	lower frequency (Hz),
f_{max}	`fMax`,	upper frequency (Hz),
.	`nF`,	number of frequencies, $\left(1 + \frac{f_{max} - f_{min}}{\delta f}\right)$,
q_M	`qM`,	mass flow rate (kg s^{-1}),
p_0	`p0`,	ambient pressure (Pa),
Q_V	`QV`,	volume flow rate (l s^{-1}),
.	`Comment`,	comment containing file history,
q_V	`qV`,	$q_V = Q_V/1000$, volume flow rate ($\mathrm{m}^3\mathrm{s}^{-1}$).

vndcr

vensa

vout

end.

B.4 End corrections

vndcr

End corrections

For $i = \{1, \cdots, N\}$,

$$
\begin{aligned}
&\text{for } E_i = 1, \quad \epsilon_{i,1} = \texttt{ec(iE,1)} &&= r_{i,2}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,1}/A_{i,2}}}{a}\right)\right)\\
&\qquad\qquad\quad \epsilon_{i,2} = \texttt{ec(iE,2)} &&= r_{i,2}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,3}/A_{i,2}}}{a}\right)\right)\\
&\qquad\qquad\quad \epsilon_{i,3} = \texttt{ec(iE,3)} &&= r_{i,4}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,3}/A_{i,4}}}{a}\right)\right)\Big|_{A_{i,4}\neq 0}\\
&\text{for } E_i = 2, \quad \epsilon_{i,j} = \texttt{ec(iE,iL)} &&= 0, \text{ for all } j\\
&\text{for } E_i = 3, \quad \epsilon_{i,j} = \texttt{ec(iE,iL)} &&= 0, \text{ for all } j\\
&\text{for } E_i = 4, \quad \epsilon_{i,1} = \texttt{ec(iE,1)} &&= -r_{i,1}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,3}/A_{i,1}}}{a}\right)\right)\\
&\qquad\qquad\quad \epsilon_{i,2} = \texttt{ec(iE,2)} &&= 2r_{i,1}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,3}/A_{i,1}}}{a}\right)\right)\\
&\qquad\qquad\quad \epsilon_{i,3} = \texttt{ec(iE,3)} &&= r_{i,5}\kappa\left(1 - \exp\left(\frac{1-\sqrt{A_{i,3}/A_{i,5}}}{a}\right)\right)\\
&\text{for } E_i = 5, \quad \epsilon_{i,j} = \texttt{ec(iE,iL)} &&= 0, \text{ for all } j.
\end{aligned}
\tag{B.4}
$$

Corrected lengths

For $i = \{1, \cdots, N\}$,

$$
\begin{aligned}
x_{i,1} &= \begin{cases} l_{i,1} - \epsilon_{i,1} & \text{for } i = 1\\ l_{i,1} - \epsilon_{i,1} + \epsilon_{i-1,3} & \text{otherwise}\end{cases}\\
x_{i,2} &= l_{i,2} + \epsilon_{i,1} + \epsilon_{i,2}\\
x_{i,3} &= \begin{cases} l_{i,3} - \epsilon_{i,2} - \epsilon_{i,3} & \text{for } E_i = 1\\ l_{i,3} & \text{otherwise}\end{cases}\\
x_{i,4} &= l_{i,4} + \epsilon_{i,3}
\end{aligned}
\tag{B.5}
$$

vensa

Define constants

Magic numbers:

$d_{1,2,3}$	=	$[0.002, 2, 0.1]$	wave number constants
$d_{4,5,6}$	=	$[0.8954, -2.146, 1.457]$	coefficients for computing R_P
$d_{7,8,9,10}$	=	$[0.0133586, -0.590789,$ $0.335762, -0.0643211]$	coefficients for computing R_0
$d_{11,12,13,14}$	=	$[0.62, 0.8, 1.0, 1.5]$	boundary values of $\frac{K_r}{K_P}$
$d_{15,16,17}$	=	$[0.1067, 1.55, -0.5417]$	coefficients for computing R
d_{18}	=	$[0.5331]$	coefficient for computing R
$d_{19,20,21}$	=	$[-2.7369, 8.4934, -4.7565]$	coefficients for computing R
$d_{22,23,24,25}$	=	$[1.03, -0.2, 0.94, 0.06]$	coefficients for computing R
$d_{26,27,28,29}$	=	$[-1.2266, 0.2336,$ $-1.2786, 0.2208]$	coefficients for computing θ
d_{30}	=	$[0.99]$	coefficient for computing $m(f)$

Frequency range:

$$f = \{f_{min}, f_{min} + \delta f, \cdots, f_{max}\} \tag{B.6}$$

Mach number, local Mach number in each atom of each element:

$$M_{i,j} = \texttt{amt(iE,iL)} = \begin{cases} 0 & \text{for } A_{i,j} = 0 \\ \dfrac{q_V}{c_0 A_{i,j}} & \text{otherwise} \end{cases} \tag{B.7}$$

Read wall parameters (from file/keyboard): Mass per unit area s_{mm}, Loss factor R_{LR}, and Natural frequency ω_0.

Wall impedance, as a function of frequency:

$$z_m(f) = \texttt{zem} = \begin{cases} 0 & \text{for } s_{mm} = 0 \\ \dfrac{\rho_0 c_0^2((\omega_0^2 - 4\pi^2 f^2)s_{mm} - j2\pi f R_{LR})}{((\omega_0^2 - 4\pi^2 f^2)^2 s_{mm}^2 + (2\pi f R_{LR})^2)} & \text{for } s_{mm} \neq 0 \end{cases} \tag{B.8}$$

Wave number:

$$k = \frac{2\pi f}{c_0} \tag{B.9}$$

Incident pressure is set to unity:

$$p_{I,0} = \texttt{pI} = 1, \qquad \text{for } f = \{f_{min}, f_{min} + \delta f, \cdots, f_{max}\} \tag{B.10}$$

B.5 Radiation

Radiation impedance (reflected pressure)

Wave number around circumference of the lips:

$$K_r = \texttt{ak} \quad = r_{1,1}k = \frac{r_{1,1}2\pi f}{c_0} \tag{B.11}$$

if isBeranek,

The equation for the reflected pressure, using Beranek's (1954) expression for a piston in an infinite baffle, is derived from the radiation impedance:

$$Z_M \quad = \quad A_{1,1}\rho_0 c_0 \left[R_1(2ka) + jX_1(2ka)\right], \tag{B.12}$$

where

$$\begin{aligned} a &= \sqrt{A_{1,1}/\pi} \\ R_1(x) &= 1 - \frac{2J_1(x)}{x} \\ X_1(x) &= \frac{4}{\pi}\left(\frac{x}{3} - \frac{x^3}{3^2.5} + \frac{x^5}{3^2.5^2.7} - \ldots\right). \end{aligned}$$

$$\Rightarrow \quad p_{R,0} \quad = \quad \frac{Z_M - 1}{Z_M + 1} \tag{B.13}$$

else,

Coefficients used for fitting impedance curve to empirical results:

$$\begin{aligned} K_P &= \texttt{akp} &&= d_1 + M_{1,1}(d_2 - M_{1,1}) - M_{1,1}^3(M_{1,1} - d_3) \\ R_P &= \texttt{rp} &&= d_4 M_{1,1} + d_5 M_{1,1}^2 + d_6 M_{1,1}^3 \\ R_0 &= \texttt{r0} &&= 1 + d_7 K_r + d_8 K_r^2 + d_9 K_r^3 + d_{10} K_r^4 \end{aligned} \tag{B.14}$$

The equation for the reflected pressure, which contains the coefficients a_n, is of the form:

$$R = \quad 1 + R_P \left[\quad a_0 \qquad\qquad +a_1\left(\tfrac{K_r}{K_P}\right) \quad +a_2\left(\tfrac{K_r}{K_P}\right)^2 \quad +a_3\left(\tfrac{K_r}{K_P}\right)^3 \right], \tag{B.15}$$

where

		$\underline{a_0}$	$\underline{a_1}$	$\underline{a_2}$	$\underline{a_3}$
$0 < \frac{K_r}{K_P}$	≤ 0.62	0	d_{15}	d_{16}	d_{17}
$0.62 < \frac{K_r}{K_P}$	≤ 0.8	$d_{18} - (d_{16}d_{11})$	d_{16}	0	0
$0.8 < \frac{K_r}{K_P}$	≤ 1.0	0	d_{19}	d_{20}	d_{21}

For higher values of K_r/K_P, R is defined

$$1.0 < \tfrac{K_r}{K_P} \quad \leq 1.5 \quad R = \mathtt{r} = (1 + R_P)[1 - 2(K_r - K_P)^2 + 3(K_r - K_P)^3]$$

$$1.5 < \tfrac{K_r}{K_P} \qquad R = \mathtt{r} = \begin{cases} R_0, & \text{for } M_{1,1} = 0 \\ R_0[1 + d_{22} M_{1,1}(1 + d_{23} M_{1,1})(d_{24} + d_{25} K_r)] & \text{otherwise} \end{cases}$$

The angle θ is calculated from the equations:

$$K_r \quad \leq 0.5 \quad \theta = \mathtt{th} = \pi + d_{26}\,(K_r) + d_{27}\,(K_r)^3\,;$$

$$0.5 < K_r \qquad \theta = \mathtt{th} = \pi + d_{28}\,(K_r) + d_{29}\,(K_r)^2\,. \tag{B.16}$$

Finally, the reflected pressure is simply the combination of the magnitude R and phase angle θ,

$$p_{R,0} = \mathtt{pR} = R \exp(j\theta), \qquad \text{for } f = \{f_{min}, f_{min} + \delta f, \cdots, f_{max}\}\,. \tag{B.17}$$

endif.

First chamber propagation transfer

Dynamic distances:

$$x_+ = \frac{x_{i,1}}{1 + M_{1,1}}, \qquad x_- = \frac{-x_{i,1}}{1 - M_{1,1}} \tag{B.18}$$

where + denotes the with-flow direction, which is the direction of progression of reflected wave (G to L), and $-$ denotes the against-flow direction, or the direction progression of incident wave (L to G).

Hydraulic reciprocal:

$$h_1 = \begin{cases} 0 & \text{for } r_{i,1} < \Delta_r \text{ (closed)} \\ \frac{1}{r_{i,1}} & \text{for } r_{i,1} \geq \Delta_r \text{ (open)} \end{cases} \tag{B.19}$$

Pressure at first junction:

$$p_{I,i} = \begin{cases} p_{I,0} \exp\left(j x_+ [k(1 + z_m(f) h_1) + \frac{\alpha}{r_{i,1}} \sqrt{f}(1 - j)]\right) & \text{for } i = 1 \\ p_{I,i-1} \exp\left(j x_+ [k(1 + z_m(f) h_1) + \frac{\alpha}{r_{i,1}} \sqrt{f}(1 - j)]\right) & \text{otherwise} \end{cases} \tag{B.20}$$

$$p_{R,i} = \begin{cases} p_{R,0} \exp\left(j x_- [k(1 + z_m(f) h_1) + \frac{\alpha}{r_{i,1}} \sqrt{f}(1 - j)]\right) & \text{for } i = 1 \\ p_{R,i-1} \exp\left(j x_- [k(1 + z_m(f)_1 h_1) + \frac{\alpha}{r_{i,1}} \sqrt{f}(1 - j)]\right) & \text{otherwise} \end{cases} \tag{B.21}$$

Work back through all elements

Call subroutine:

for $i = 1$ to N,
    if $E_i == 1$ then,
        etype1;
    elseif $E_i == 2$ then,
        etype2;
    elseif $E_i == 3$ then,
        etype3;
    elseif $E_i == 4$ then,
        etype4;
    elseif $E_i \neq 5$ then,
        error
    endif
endfor

B.6 Element transfers

B.6.1 Orifice

etype1

Pressure transfer (inlet)

$$
\begin{aligned}
m(f) &= d_{30}\exp\left(\frac{-2\alpha\epsilon_{i,1}\sqrt{f}}{(r_{i,1}-r_{i,2})}\right)\\
ga(f) &= (1+M_{i,2})(M_{i,2}(\gamma-1)+1)+\left(\frac{A_{i,1}-A_{i,2}}{A_{i,2}}\right)\frac{m(f)-\exp(j2\epsilon_{i,1}k)}{m(f)+\exp(j2\epsilon_{i,1}k)}\\
gb(f) &= (1-M_{i,2})(M_{i,2}(\gamma-1)-1)+\left(\frac{A_{i,1}-A_{i,2}}{A_{i,2}}\right)\frac{m(f)-\exp(j2\epsilon_{i,1}k)}{m(f)+\exp(j2\epsilon_{i,1}k)}\\
gc(f) &= M_{i,2}(\gamma-1)\left((1+M_{i,2})+\left(\frac{p_{R,i}(f)}{p_{I,i}(f)}\right)(1-M_{i,2})\right)\\
&\qquad +\frac{A_{i,1}}{A_{i,2}}\left(\left(1+M_{i,2}\frac{A_{i,2}}{A_{i,1}}\right)-\left(\frac{p_{R,i}(f)}{p_{I,i}(f)}\right)\left(1-M_{i,2}\frac{A_{i,2}}{A_{i,1}}\right)\right)\\
&= \left[(1+M_{i,2})(\gamma-1)M_{i,2}+M_{i,2}+\frac{A_{i,1}}{A_{i,2}}\right]\\
&\qquad +\left[(1-M_{i,2})(\gamma-1)M_{i,2}+M_{i,2}-\frac{A_{i,1}}{A_{i,2}}\right]\frac{p_{R,i}(f)}{p_{I,i}(f)}\\
&= \left(\frac{p_{R,i}(f)}{p_{I,i}(f)}\right)\left[(\gamma-M_{i,2}\gamma+M_{i,2})M_{i,2}-\frac{A_{i,1}}{A_{i,2}}\right]\\
&\qquad +\left[(\gamma+M_{i,2}\gamma-M_{i,2})M_{i,2}+\frac{A_{i,1}}{A_{i,2}}\right]\\
gg(f) &= \frac{A_{i,1}}{A_{i,2}}+M_{i,2}-M_{i,2}\left(\frac{A_{i,1}-A_{i,2}}{A_{i,2}}\right)\frac{m(f)-\exp(j2\epsilon_{i,1}k)}{m(f)+\exp(j2\epsilon_{i,1}k)}\\
gh(f) &= \frac{A_{i,1}}{A_{i,2}}-M_{i,2}-M_{i,2}\left(\frac{A_{i,1}-A_{i,2}}{A_{i,2}}\right)\frac{m(f)-\exp(j2\epsilon_{i,1}k)}{m(f)+\exp(j2\epsilon_{i,1}k)}\\
gj(f) &= \left[\frac{A_{i,1}}{A_{i,2}}(1-M_{i,2})+2M_{i,2}\right]+\frac{p_{R,i}(f)}{p_{I,i}(f)}\left[\frac{A_{i,1}}{A_{i,2}}(1+M_{i,2})-2M_{i,2}\right]
\end{aligned}
$$

$$
\begin{aligned}
qp_{I,1}(f) &= p_{I,i}\frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)}\\
qp_{R,1}(f) &= \frac{p_{I,i}}{gb}\left(gc-ga\frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)}\right)
\end{aligned}
\tag{B.22}
$$

Second propagation transfer

Dynamic distances:

$$
x_+=\frac{x_{i,2}}{1+M_{i,2}},\qquad x_-=\frac{-x_{i,2}}{1-M_{i,2}} \tag{B.23}
$$

Hydraulic reciprocal:

$$h_2 = \begin{cases} 0 & \text{for } r_{i,2} < \Delta_r \text{ (closed)} \\ \frac{1}{r_{i,2}} & \text{for } r_{i,2} \geq \Delta_r \text{ (open)} \end{cases} \tag{B.24}$$

Elemental pressure (delay):

$$\begin{aligned} qp_{I,2}(f) &= qp_{I,1}(f) \exp\left(jx_+ \left[k(1 + z_m(f)h_2) + \tfrac{\alpha\sqrt{f}}{r_{i,2}}(1-j)\right]\right) \\ qp_{R,2}(f) &= qp_{R,1}(f) \exp\left(jx_- \left[k(1 + z_m(f)h_2) + \tfrac{\alpha\sqrt{f}}{r_{i,2}}(1-j)\right]\right) \end{aligned} \tag{B.25}$$

Elemental pressure transfer (expansion)

if expansion$\{r_{i,3} \geq (r_{i,2} + \Delta_0)\}$ then,

Note: should test $A_{i,3} \geq A_{i,2} + \Delta_0$!

$$\begin{aligned} m(f) &= d_{30} \exp\left(\frac{-2\alpha\epsilon_{i,2}\sqrt{f}}{(r_{i,3} - r_{i,2})}\right) \\ gg(f) &= \left(1 + M_{i,2}\frac{A_{i,2}}{A_{i,3}}\right) - \left(\frac{A_{i,3} - A_{i,2}}{A_{i,3}}\right)\frac{1 - m(f)\exp(-j2\epsilon_{i,2}k)}{1 + m(f)\exp(-j2\epsilon_{i,2}k)} \\ gh(f) &= \left(1 - M_{i,2}\frac{A_{i,2}}{A_{i,3}}\right) + \left(\frac{A_{i,3} - A_{i,2}}{A_{i,3}}\right)\frac{1 - m(f)\exp(-j2\epsilon_{i,2}k)}{1 + m(f)\exp(-j2\epsilon_{i,2}k)} \\ gj(f) &= \frac{A_{i,2}}{A_{i,3}}\left(1 + M_{i,2} - (1 - M_{i,2})\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right) \\ ga(f) &= \left(1 + M_{i,2} + (1 - M_{i,2})\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right) \\ qp_{I,3}(f) &= qp_{I,2}\frac{(ga.gh + gj.(1 - M_{i,2}))}{(gh.(1 + M_{i,2}) + gg.(1 - M_{i,2}))} \\ qp_{R,3}(f) &= \frac{[qp_{I,2}(f).ga - qp_{I,3}(f)(1 + M_{i,2})]}{(1 - M_{i,2})} \end{aligned} \tag{B.26}$$

Third propagation transfer

Dynamic distances:

if element continues$\{x_{i,3} \geq \Delta_0\}$ then,

$$x_+ = \frac{x_{i,3}}{1 + M_{i,3}}, \qquad x_- = \frac{-x_{i,3}}{1 - M_{i,3}} \tag{B.27}$$

Hydraulic reciprocal:

$$h_3 = \begin{cases} 0 & \text{for } r_{i,3} < \Delta_r \text{ (closed)} \\ \frac{1}{r_{i,3}} & \text{for } r_{i,3} \geq \Delta_r \text{ (open)} \end{cases} \tag{B.28}$$

Elemental pressure (delay):

$$\begin{aligned} qp_{I,4}(f) &= qp_{I,3}(f) \exp\left(jx_+ \left[k(1 + z_m(f)h_3) + \tfrac{\alpha\sqrt{f}}{r_{i,3}}(1-j)\right]\right) \\ qp_{R,4}(f) &= qp_{R,3}(f) \exp\left(jx_- \left[k(1 + z_m(f)h_3) + \tfrac{\alpha\sqrt{f}}{r_{i,3}}(1-j)\right]\right) \end{aligned} \tag{B.29}$$

Elemental pressure transfer (inlet)

if expansion$\{r_{i,3} \geq (r_{i,4} + \Delta_0)\}$ then,

$$
\begin{aligned}
m(f) &= d_{30} \exp\left(\frac{-2\alpha\epsilon_{i,3}\sqrt{f}}{(r_{i,3} - r_{i,4})}\right) \\
ga(f) &= (1 + M_{i,4})(M_{i,4}(\gamma - 1) + 1) + \left(\frac{A_{i,3} - A_{i,4}}{A_{i,4}}\right)\frac{m(f) - \exp(j2\epsilon_{i,3}k)}{m(f) + \exp(j2\epsilon_{i,3}k)} \\
gb(f) &= (1 - M_{i,4})(M_{i,4}(\gamma - 1) - 1) + \left(\frac{A_{i,3} - A_{i,4}}{A_{i,4}}\right)\frac{m(f) - \exp(j2\epsilon_{i,3}k)}{m(f) + \exp(j2\epsilon_{i,3}k)} \\
gc(f) &= M_{i,4}(\gamma - 1)\left((1 + M_{i,4}) + \left(\frac{qp_{R,4}(f)}{qp_{I,4}(f)}\right)(1 - M_{i,4})\right) \\
&\qquad + \frac{A_{i,3}}{A_{i,4}}\left(\left(1 + M_{i,4}\frac{A_{i,4}}{A_{i,3}}\right) - \left(\frac{qp_{R,i}(f)}{qp_{I,i}(f)}\right)\left(1 - M_{i,4}\frac{A_{i,4}}{A_{i,3}}\right)\right) \\
&= \left(\frac{qp_{R,4}(f)}{qp_{I,4}(f)}\right)\left[(1 - M_{i,4})(\gamma - 1)M_{i,4} + M_{i,4} - \frac{A_{i,3}}{A_{i,4}}\right] \\
&\qquad + \left[(1 + M_{i,4})(\gamma - 1)M_{i,4} + M_{i,4} + \frac{A_{i,3}}{A_{i,4}}\right] \\
&= \left(\frac{qp_{R,4}(f)}{qp_{I,4}(f)}\right)\left[(\gamma - M_{i,4}\gamma + M_{i,4})M_{i,4} - \frac{A_{i,3}}{A_{i,4}}\right] \\
&\qquad + \left[(\gamma + M_{i,4}\gamma - M_{i,4})M_{i,4} + \frac{A_{i,3}}{A_{i,4}}\right] \\
gg(f) &= \frac{A_{i,3}}{A_{i,4}} + M_{i,4} - M_{i,4}\left(\frac{A_{i,3} - A_{i,4}}{A_{i,4}}\right)\frac{m(f) - \exp(j2\epsilon_{i,3}k)}{m(f) + \exp(j2\epsilon_{i,3}k)} \\
gh(f) &= \frac{A_{i,3}}{A_{i,4}} - M_{i,4} - M_{i,4}\left(\frac{A_{i,3} - A_{i,4}}{A_{i,4}}\right)\frac{m(f) - \exp(j2\epsilon_{i,3}k)}{m(f) + \exp(j2\epsilon_{i,3}k)} \\
gj(f) &= \left[\frac{A_{i,3}}{A_{i,4}}(1 - M_{i,4}) + 2M_{i,4}\right] + \frac{p_{R,i}(f)}{p_{I,i}(f)}\left[\frac{A_{i,3}}{A_{i,4}}(1 + M_{i,4}) - 2M_{i,4}\right] \\
p_{I,1}(f) &= qp_{I,4}\frac{(gc.gh - gj.gb)}{(ga.gh - gg.gb)} \\
p_{R,1}(f) &= \frac{qp_{I,4}}{gb}\left(gc - ga\frac{(gc.gh - gj.gb)}{(ga.gh - gg.gb)}\right)
\end{aligned}
\tag{B.30}
$$

else,

$$p_{I,1}(f) = qp_{I,4}$$
$$p_{R,1}(f) = qp_{R,4}$$

endif $\{r_{i,3} \geq (r_{i,4} + \Delta_0)\}$

else,

$$p_{I,1}(f) = qp_{I,3}$$
$$p_{R,1}(f) = qp_{R,3}$$

endif $\{x_{i,3} \geq \Delta_0\}$

else,

$$p_{I,1}(f) = qp_{I,2}$$
$$p_{R,1}(f) = qp_{R,2}$$

endif $\{r_{i,3} \geq (r_{i,2} + \Delta_0)\}$

return.

B.6.2 RAMP

etype2

Pressure transfer (ramp)

Dynamic distances:

$$x_+ = \frac{x_{i,2}}{1+\left(\frac{M_{i,1}+M_{i,2}}{2}\right)}, \qquad x_- = \frac{x_{i,2}}{1-\left(\frac{M_{i,1}+M_{i,2}}{2}\right)} \tag{B.31}$$

Hydraulic reciprocal:

$$\bar{r} = \frac{(r_{i,1} + r_{i,2})}{2} \tag{B.32}$$

$$h_1 = \begin{cases} 0 & \text{for } \bar{r} < \Delta_r \text{ (closed)} \\ \frac{1}{\bar{r}} & \text{for } \bar{r} \geq \Delta_r \text{ (open)} \end{cases} \tag{B.33}$$

Elemental pressure:

$$\begin{aligned} qp_{I,1}(f) &= p_{I,i}(f) \exp\left(jx_+ \left[k(1 + z_m(f)h_1) + \frac{\alpha\sqrt{f}}{\bar{r}}(1-j)\right]\right) \\ qp_{R,1}(f) &= p_{R,i}(f) \exp\left(jx_- \left[k(1 + z_m(f)h_1) + \frac{\alpha\sqrt{f}}{\bar{r}}(1-j)\right]\right) \end{aligned} \tag{B.34}$$

$$\begin{aligned} p_{I,i}(f) &= \frac{(1+M_{i,1})}{(1+M_{i,2})}\frac{(A_{i,2}+A_{i,1})}{2A_{i,2}}qp_{I,1}(f) + \frac{(1-M_{i,1})}{(1+M_{i,2})}\frac{(A_{i,2}-A_{i,1})}{2A_{i,2}}qp_{R,1}(f) \\ p_{R,i}(f) &= \frac{(1+M_{i,1})}{(1-M_{i,2})}\frac{(A_{i,2}-A_{i,1})}{2A_{i,2}}qp_{I,1}(f) + \frac{(1-M_{i,1})}{(1-M_{i,2})}\frac{(A_{i,2}+A_{i,1})}{2A_{i,2}}qp_{R,1}(f) \end{aligned} \tag{B.35}$$

return.

B.6.3 CONE

etype3

Pressure transfer (cone)

Cone angle, hydraulic reciprocal and quadratic constant:

$$\phi = \arctan\frac{(r_{i,1}-r_{i,2})}{l_{i,2}}, \qquad h_1 = \frac{\sin\phi}{r_{i,1}}, \qquad ? = \frac{3(1-\cos\phi)(2+\cos\phi)}{(3+\cos\phi)} \tag{B.36}$$

Elemental pressure:

$$\begin{aligned}
ga(\phi) &= \frac{1}{h_1}\left(h_1^2 + jkh_1 - k^2?\right)\\
gb(\phi) &= \frac{1}{h_1}\left(h_1^2 - jkh_1 - k^2?\right)\\
gc(f,\phi) &= \frac{1}{\rho c_0}\left[p_{I,i}(f)\left(1+M_{i,1}-j\frac{k?}{h_1}\right) - p_{R,i}(f)\left(1-M_{i,1}+j\frac{k?}{h_1}\right)\right]\\
gg(\phi) &= M_{i,1}h_1 - \frac{k^2?}{h_1} + jk\left(1+M_{i,1}+\frac{?}{h_1}\right)\\
gh(\phi) &= M_{i,1}h_1 + \frac{k^2?}{h_1} + jk\left(1-M_{i,1}+\frac{?}{h_1}\right)\\
gj(f,\phi) &= \frac{1}{\rho c_0}\left[p_{I,i}(f)\left(1+2M_{i,1}-j\frac{k?}{h_1}\right) - p_{R,i}(f)\left(1-2M_{i,1}+j\frac{k?}{h_1}\right)\right]
\end{aligned}$$

$$\begin{aligned}
qp_{I,1}(f) &= \frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)}\frac{r_{i,1}}{r_{i,2}}\exp\left(\frac{jk}{\sin\phi}\left[(1+M_{i,1})r_{i,1}-(1+M_{i,2})r_{i,2}\right]\right)\\
qp_{R,1}(f) &= \left(gc - ga\frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)}\right)\frac{r_{i,1}}{r_{i,2}}\exp\left(\frac{jk}{\sin\phi}\left[(1-M_{i,2})r_{i,2}-(1-M_{i,1})r_{i,1}\right]\right)
\end{aligned} \tag{B.37}$$

Conical pressure transfer

Hydraulic reciprocal:

$$h_2 = \frac{\sin\phi}{r_{i,2}} \tag{B.38}$$

Elemental pressure:

$$\begin{aligned}
ga(\phi) &= 1 + M_{i,2} - j\frac{k?}{h_2} \\
gb(\phi) &= 1 - M_{i,2} + j\frac{k?}{h_2} \\
gc(f,\phi) &= \rho c_0 \left[qp_{I,1}(f) \left(\frac{1}{h_2} \left(h_2^2 + jkh_2 - k^2? \right) \right) \right. \\
&\qquad \left. + qp_{R,1}(f) \left(\frac{1}{h_2} \left(h_2^2 - jkh_2 - k^2? \right) \right) \right] \\
gg(\phi) &= 1 + 2M_{i,12} - j\frac{k?}{h_2} \\
gh(\phi) &= 1 - 2M_{i,2} + j\frac{k?}{h_2} \\
gj(f,\phi) &= \rho c_0 \left[qp_{I,1}(f) \left(M_{i,2}h_2 - \frac{k^2?}{h_2} + jk \left(1 + M_{i,2} + \frac{?}{h_2} \right) \right) \right. \\
&\qquad \left. - qp_{R,1}(f) \left(M_{i,2}h_2 + \frac{k^2?}{h_2} + jk \left(1 - M_{i,2} + \frac{?}{h_2} \right) \right) \right] \\
p_{I,1}(f) &= \frac{(gc.gh + gj.gb)}{(ga.gh + gg.gb)} \\
p_{R,1}(f) &= \frac{1}{gb} \left(ga \frac{(gc.gh - gj.gb)}{(ga.gh + gg.gb)} - gc \right)
\end{aligned} \tag{B.39}$$

return.

B.6.4 OUTLET

etype4

Pressure transfer (expansion)

if expansion$\{A_{i,3} \geq (A_{i,1} + \Delta_0)\}$ then,

$$\bar{r} = \begin{cases} r_{i,2} & \text{for } r_{i,2} \neq 0 \\ r_{i,3} - r_{i,1} & \text{otherwise} \end{cases} \tag{B.40}$$

Hydraulic reciprocal:

$$h_1 = \begin{cases} \frac{1}{\bar{r}} & \text{for } \bar{r} \geq \Delta_0 \\ 0 & \text{otherwise} \end{cases} \tag{B.41}$$

if forward sinus$\{A_{i,2} > \Delta_0\}$ then,

Outlet reflection coefficient and elemental pressure:

$$\begin{aligned}
m(f) &= \begin{cases} d_{30} \exp\left(\dfrac{-2\alpha x_{i,2}\sqrt{f}}{\bar{r}}\right) & \text{for } x_{i,2} > \Delta_0 \\ d_{30} & \text{otherwise} \end{cases} \\
gg(f) &= \left(1 + M_{i,1}\frac{A_{i,1}}{A_{i,3}}\right) - \left(\frac{A_{i,2}}{A_{i,3}}\right)\frac{1 - m(f)\exp\left(-j2x_{i,2}k(1 + z_m(f)h_1)\right)}{1 + m(f)\exp\left(-j2x_{i,2}k(1 + z_m(f)h_1)\right)} \\
gh(f) &= \left(1 - M_{i,1}\frac{A_{i,1}}{A_{i,3}}\right) + \left(\frac{A_{i,2}}{A_{i,3}}\right)\frac{1 - m(f)\exp\left(-j2x_{i,2}k(1 + z_m(f)h_1)\right)}{1 + m(f)\exp\left(-j2x_{i,2}k(1 + z_m(f)h_1)\right)} \\
gj(f) &= \frac{A_{i,1}}{A_{i,3}}\left(1 + M_{i,1} - (1 - M_{i,1})\frac{p_{R,i}(f)}{p_{I,i}(f)}\right) \\
ga(f) &= \left(1 + M_{i,1} + (1 - M_{i,1})\frac{p_{R,i}(f)}{p_{I,i}(f)}\right) \\
qp_{I,1}(f) &= p_{I,i}(f)\frac{(ga.gh + gj.(1 - M_{i,1}))}{(gh.(1 + M_{i,1}) + gg.(1 - M_{i,1}))} \\
qp_{R,1}(f) &= \frac{[p_{I,i}(f).ga - qp_{I,1}(f)(1 + M_{i,1})]}{(1 - M_{i,1})}
\end{aligned} \tag{B.42}$$

else,

Outlet reflection coefficient and elemental pressure:

$$\begin{aligned}
m(f) &= \begin{cases} d_{30} \exp\left(\dfrac{-2\alpha x_{i,2}\sqrt{f}}{\bar{r}}\right) & \text{for } x_{i,2} > \Delta_0 \\ d_{30} & \text{otherwise} \end{cases} \\
gg(f) &= \left(1 + M_{i,1}\frac{A_{i,1}}{A_{i,3}}\right) - \left(\frac{A_{i,3} - A_{i,1}}{A_{i,3}}\right)\frac{1 - m(f)\exp\left(-j2x_{i,2}k\right)}{1 + m(f)\exp\left(-j2x_{i,2}k\right)} \\
gh(f) &= \left(1 - M_{i,1}\frac{A_{i,1}}{A_{i,3}}\right) + \left(\frac{A_{i,3} - A_{i,1}}{A_{i,3}}\right)\frac{1 - m(f)\exp\left(-j2x_{i,2}k\right)}{1 + m(f)\exp\left(-j2x_{i,2}k\right)}
\end{aligned}$$

$$
\begin{aligned}
gj(f) &= \frac{A_{i,1}}{A_{i,3}}\left(1+M_{i,1}-(1-M_{i,1})\frac{p_{R,i}(f)}{p_{I,i}(f)}\right)\\
ga(f) &= \left(1+M_{i,1}+(1-M_{i,1})\frac{p_{R,i}(f)}{p_{I,i}(f)}\right)\\
qp_{I,1}(f) &= p_{I,i}(f)\frac{(ga.gh+gj.(1-M_{i,1}))}{(gh.(1+M_{i,1})+gg.(1-M_{i,1}))}\\
qp_{R,1}(f) &= \frac{[p_{I,i}(f).ga-qp_{I,1}(f)(1+M_{i,1})]}{(1-M_{i,1})}
\end{aligned}
\tag{B.43}
$$

endif.

else,

$$
\begin{aligned}
qp_{I,1}(f) &= p_{I,i}(f)\\
qp_{R,1}(f) &= p_{R,i}(f)
\end{aligned}
\tag{B.44}
$$

endif

Second propagation transfer

Dynamic distances:

$$
x_+ = \frac{(x_{i,3}-x_{i,2}-x_{i,4})}{1+M_{i,3}}, \qquad x_- = \frac{-(x_{i,3}-x_{i,2}-x_{i,4})}{1-M_{i,3}} \tag{B.45}
$$

Hydraulic reciprocal:

$$
h_2 = \begin{cases} 0 & \text{for } r_{i,3} < \Delta_r \text{ (closed)} \\ \frac{1}{r_{i,3}} & \text{for } r_{i,3} \geq \Delta_r \text{ (open)} \end{cases} \tag{B.46}
$$

Elemental pressure:

$$
\begin{aligned}
qp_{I,2}(f) &= qp_{I,1}(f)\exp\left(jx_+\left[k(1+z_m(f)h_2)+\tfrac{\alpha\sqrt{f}}{r_{i,3}}(1-j)\right]\right)\\
qp_{R,2}(f) &= qp_{R,1}(f)\exp\left(jx_-\left[k(1+z_m(f)h_2)+\tfrac{\alpha\sqrt{f}}{r_{i,3}}(1-j)\right]\right)
\end{aligned}
\tag{B.47}
$$

Second propagation transfer

if contraction$\{A_{i,3} \geq (A_{i,5}+\Delta_0)\}$ then,

Hydraulic reciprocal:

$$
\begin{aligned}
\bar{r} &= \begin{cases} r_{i,4} & \text{for } r_{i,4} \neq 0 \\ r_{i,3}-r_{i,5} & \text{otherwise} \end{cases}\\
h_3 &= \begin{cases} \frac{1}{\bar{r}} & \text{for } \bar{r} \geq \Delta_r \text{ (open)} \\ 0 & \text{otherwise (closed)} \end{cases}
\end{aligned}
\tag{B.48}
$$

if backward sinus$\{A_{i,4} > \Delta_0\}$ then,

Elemental pressure (inlet):

$$
\begin{aligned}
m(f) &= \begin{cases} d_{30}\exp\left(\dfrac{-2\alpha x_{i,4}\sqrt{f}}{\bar{r}}\right) & \text{for } x_{i,4} < \Delta_0 \text{ (closed)} \\ d_{30} & \text{for } x_{i,4} \geq \Delta_0 \text{ (open} \end{cases} \\
ga(f) &= (1+M_{i,5})(M_{i,5}(\gamma-1)+1) + \left(\frac{A_{i,4}}{A_{i,3}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)}{m(f)+\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)} \\
gb(f) &= (1-M_{i,5})(M_{i,5}(\gamma-1)-1) + \left(\frac{A_{i,4}}{A_{i,3}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)}{m(f)+\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)} \\
gc(f) &= M_{i,5}(\gamma-1)\left((1+M_{i,5}) + \left(\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right)(1-M_{i,5})\right) \\
gg(f) &= \frac{A_{i,3}}{A_{i,5}} + M_{i,5} - M_{i,5}\left(\frac{A_{i,4}}{A_{i,3}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)}{m(f)+\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)} \\
gh(f) &= \frac{A_{i,3}}{A_{i,5}} - M_{i,5} - M_{i,5}\left(\frac{A_{i,4}}{A_{i,3}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)}{m(f)+\exp\left(j2x_{i,4}k(1+z_m(f)h_3)\right)} \\
gj(f) &= \left(\frac{A_{i,3}}{A_{i,5}}(1-M_{i,5})+2M_{i,5}\right) + \left(\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right)\left(\frac{A_{i,3}}{A_{i,5}}(1+M_{i,5})-2M_{i,5}\right) \\
p_{I,i}(f) &= qp_{I,2}(f)\frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)} \\
p_{R,i}(f) &= \frac{qp_{I,2}(f)}{gb}\left(gc - ga\frac{(gc.gh-gj.gb)}{(ga.gh-gg.gb)}\right)
\end{aligned}
\tag{B.49}
$$

else,

Elemental pressure (inlet):

$$
\begin{aligned}
m(f) &= \begin{cases} d_{30}\exp\left(\dfrac{-2\alpha x_{i,4}\sqrt{f}}{\bar{r}}\right) & \text{for } x_{i,4} < \Delta_0 \text{ (closed)} \\ d_{30} & \text{for } x_{i,4} \geq \Delta_0 \text{ (open} \end{cases} \\
ga(f) &= (1+M_{i,5})(M_{i,5}(\gamma-1)+1) + \left(\frac{A_{i,3}-A_{i,5}}{A_{i,5}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k\right)}{m(f)+\exp\left(j2x_{i,4}k\right)} \\
gb(f) &= (1-M_{i,5})(M_{i,5}(\gamma-1)-1) + \left(\frac{A_{i,3}-A_{i,5}}{A_{i,5}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k\right)}{m(f)+\exp\left(j2x_{i,4}k\right)} \\
gc(f) &= M_{i,5}(\gamma-1)\left((1+M_{i,5}) + \left(\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right)(1-M_{i,5})\right) \\
gg(f) &= \frac{A_{i,3}}{A_{i,5}} + M_{i,5} - M_{i,5}\left(\frac{A_{i,3}-A_{i,5}}{A_{i,5}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k\right)}{m(f)+\exp\left(j2x_{i,4}k\right)} \\
gh(f) &= \frac{A_{i,3}}{A_{i,5}} - M_{i,5} - M_{i,5}\left(\frac{A_{i,3}-A_{i,5}}{A_{i,5}}\right)\frac{m(f)-\exp\left(j2x_{i,4}k\right)}{m(f)+\exp\left(j2x_{i,4}k\right)} \\
gj(f) &= \left(\frac{A_{i,3}}{A_{i,5}}(1-M_{i,5})+2M_{i,5}\right) + \left(\frac{qp_{R,2}(f)}{qp_{I,2}(f)}\right)\left(\frac{A_{i,3}}{A_{i,5}}(1+M_{i,5})-2M_{i,5}\right)
\end{aligned}
$$

$$\begin{aligned} p_{I,i}(f) &= qp_{I,2}(f)\frac{(gc.gh - gj.gb)}{(ga.gh - gg.gb)} \\ p_{R,i}(f) &= \frac{qp_{I,2}(f)}{gb}\left(gc - ga\frac{(gc.gh - gj.gb)}{(ga.gh - gg.gb)}\right) \end{aligned} \tag{B.50}$$

endif.

else,

$$\begin{aligned} p_{I,i}(f) &= qp_{I,2}(f) \\ p_{R,i}(f) &= qp_{R,2}(f) \end{aligned} \tag{B.51}$$

endif

return.

B.7 Outputs

vout

B.7.1 Glottal quantities

Glottal reflection coefficient:

$$R_G = \texttt{RG} = \frac{p_{R,N}(f)}{p_{I,N}(f)} \tag{B.52}$$

Driving-point impedance at glottis:

$$z_G = \texttt{zG} = \frac{p_G}{u_G} = \rho c_0 \frac{p_{I,N}(f) + p_{R,N}(f)}{p_{I,N}(f) - p_{R,N}(f)} \tag{B.53}$$

Glottal pressure:

$$p_G = \texttt{pG} = p_{R,N}(f) + p_{I,N}(f) \tag{B.54}$$

Glottal acoustic velocity:

$$u_G = \texttt{uG} = \frac{p_{I,N}(f) - p_{R,N}(f)}{\rho c_0} \tag{B.55}$$

Glottal force:

$$F_G = \texttt{FG} = jk\left(p_{I,N}(f) - p_{R,N}(f)\right) \tag{B.56}$$

B.7.2 Losses

Radiation impedance:

$$z_{rad}(f) = \texttt{zRad} = \frac{p_L(f)}{U_L(f)} = \frac{p_L(f)}{u_L(f)A_L} = \frac{\rho_0 c_0 \left(p_{I,0}(f) + p_{R,0}(f)\right)}{A_L \left(p_{I,0}(f) - p_{R,0}(f)\right)} \tag{B.57}$$

Attenuation:

$$H(f) = \texttt{H} = \frac{p_{I,G}}{p_{I,L}} = \frac{p_{I,N}(f)}{p_{I,0}(f)} \tag{B.58}$$

Wall impedance:

$$z_m(f) = \texttt{zem} \tag{B.59}$$

B.7.3 Transfer functions

Volume-velocity VTTF:

$$H^V(f) = \texttt{HV} = \frac{U_L(f)}{U_G(f)} = \frac{u_L(f)A_L}{u_G(f)A_G} = \frac{A_L}{A_G}\frac{p_{I,0}(f) - p_{R,0}(f)}{p_{I,N}(f) - p_{R,N}(f)} \tag{B.60}$$

Pressure VTTF:

$$H^P(f) = \texttt{HP} = \frac{p_L(f)}{U_G(f)} = \frac{p_L(f)}{u_G(f)A_G} = \frac{\rho_0 c_0 \left(p_{I,0}(f) + p_{R,0}(f)\right)}{A_G \left(p_{I,N}(f) - p_{R,N}(f)\right)} \tag{B.61}$$

return.

Appendix C

Vocal-tract dimensions

C.1 Basic physiology

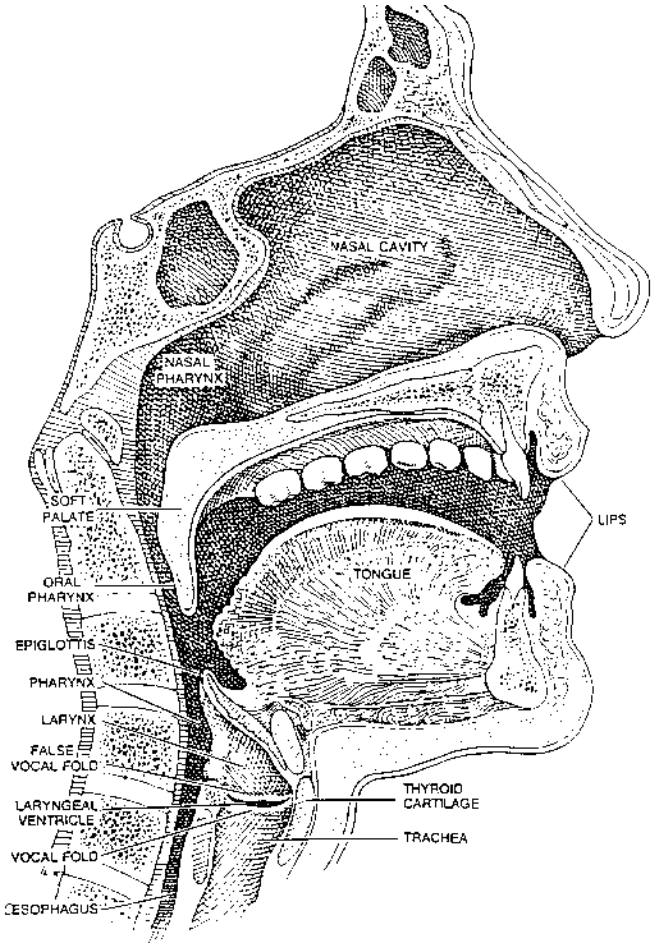


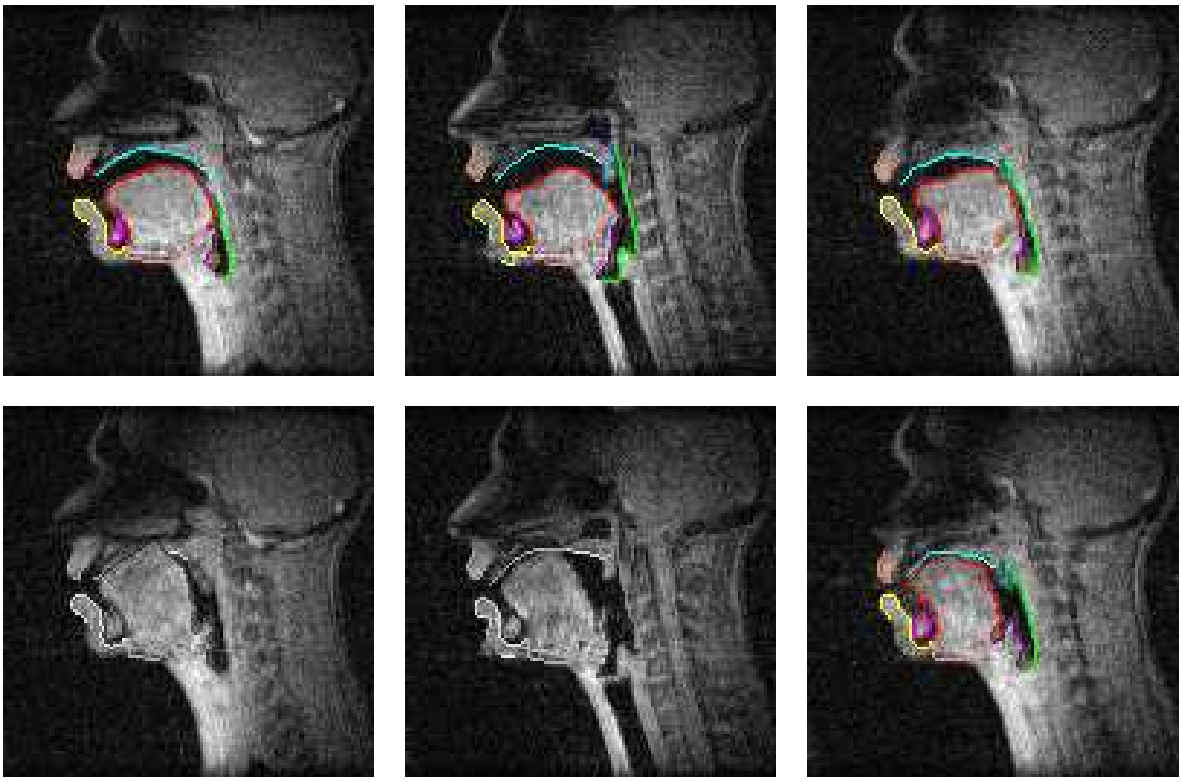
Figure C.1: Sagittal, or longitudinal, section of the human vocal apparatus, reprinted from Sundberg (1977).

Figure C.1 is a sketch taken from Sundberg (1977), which shows the airways that are involved in sound production during speech, except the lungs which are connected via the trachea: the

larynx, the pharynx, and the oral and nasal cavities, which together constitute the vocal tract. The shape of these passageways is modified by the tongue, the lips, the jaw and the velum, which hangs down from the soft palate. The epiglottis covers the larynx during swallowing to prevent any unwanted food stuffs from entering the trachea, but is normally held open during speech and lies close to the back of the tongue. These parts of anatomy are often referred to as the articulators since the adjustment of the geometry along the vocal tract allows for the full range of sounds that make up our phonetic repertory to be produced.

C.2 Vocal-tract outlines

Figure C.2 gives the dynamic MRI frames for the two vowels in [pʰɑsi] with each section of the vocal-tract outline marked with a certain grey level. Mohammad's thesis (1999) describes how the dMRI pictures were captured. The left, right and mid-sagittal outlines for each of the four phonemes, [p], [ɑ], [s] and [i], are given in Figure C.3. Details of their interpretation and conversion to area functions are given in Chapter 3 of the present thesis.



Left Mid-sagittal Right

Figure C.2: Sagittal dMRI slices, left, middle and right, for the vowels in [pʰɑsi] by PJ: (top) [ɑ] and (bottom) [i], frames 10 and 31 respectively. The segmented outlines are overlaid in various shades of grey, and include the lower mandible but not the teeth.

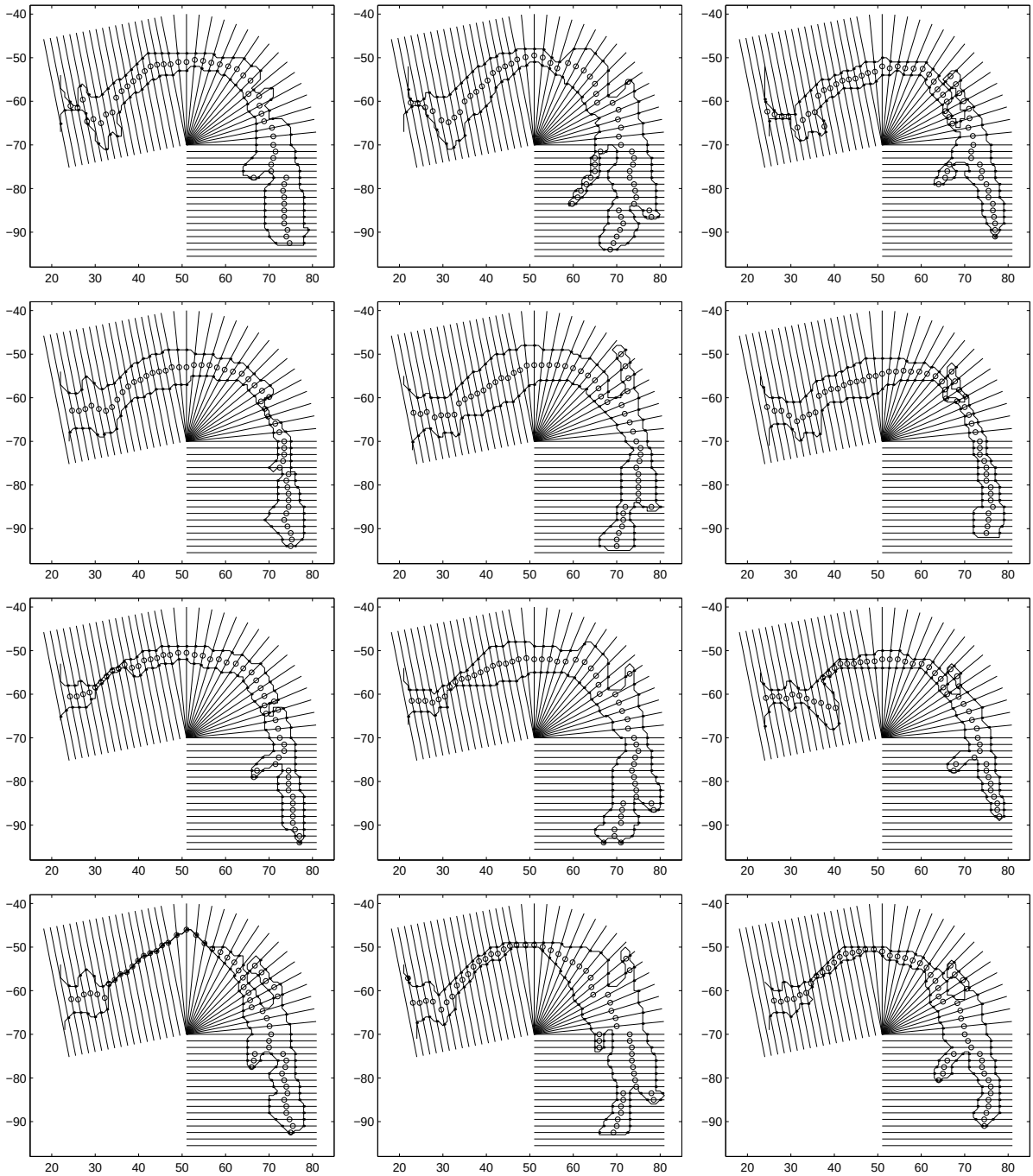


Figure C.3: Outlines (thick) from the three sagittal dMRI slices (left, middle and right) for each phoneme in [p$^{\text{h}}$ɑsi] spoken by PJ: (from top) [p], [ɑ], [s] and [i]. The x-axis is the x coordinate, and the y-axis the y coordinate. Grid lines (thin) are superimposed and the intercepts are marked by dots, and circles indicate the mid-point of each vocal-tract section.

Appendix D

Periodic-aperiodic decomposition

D.1 Introduction

The periodic-aperiodic decomposition (PAPD) is an alternative to the PSHF technique described in this thesis. It was developed by Yegnanarayana, d'Alessandro, and Darsinos (1998) for separating the voiced and unvoiced components of a mixed-source speech signal. The algorithm would appear to have the characteristics needed for our purposes, and indeed we have adopted aspects of their general approach. However, as mentioned in Chapter 5, we have discovered certain problems with it, which we have used to inform the development of our PSHF. This critique summarises their algorithm, argues that the interpolation procedure converges to the original signal, presents supporting simulation results and discusses their approach in general. Although the method was first published at conferences (d'Alessandro et al. 1995; Darsinos et al. 1995), our treatment of the technique concentrates on the more substantial journal publications (Yegnanarayana et al. 1998; d'Alessandro et al. 1998). For consistency of notation within this thesis, many of their symbols have been altered. The substitutions are given in Table D.1. A lengthy quotation using our symbols is appended for ease of reference and for completeness.

Publications using the PAPD give some results of application to synthetic and recorded speech examples, a summary of which is given in Table D.2. Their results of decomposing synthetic signals show a strong correlation between the HNR that was prescribed when generating the synthetic speech (prescribed HNR), and the value calculated from the decomposed signals (measured HNR), which they called the periodic-aperiodic energy ratio. However, there appears to be a tendency to under-estimate the aperiodic component, since all reported values of measured HNR were too high, except in the total absence of noise. The effects of jitter, shimmer and f_0 glides are also highly significant, producing a large reduction in the measured HNR; a normal degree of jitter ($\sim 1\,\%$) typically gives errors of the order of $10\,\%$ on the periodic component (i.e., HNR ≈ 20 dB).

Here	YAD	Description
n	n	point in time
k	k	point in frequency
N	N	DFT length
m	m	iteration number
B_d	F_r	selected aperiodic bins
$a(n)$	$e(n)$	excitation signal
$A_w(k)$	$E(k)$	excitation spectrum
$d(n)$	$r(n)$	true aperiodic excitation signal
$D(k)$	$R(k)$	true aperiodic part's spectrum
$G(k)$	$P(k)$	true periodic part's spectrum
$d_0(n)$	$r_0(n)$	initial estimate of aperiodic signal
$D_0(k)$	$R_0(k)$	initial estimate of aperiodic spectrum
$'$	$\hat{}$	time-compacted version
$d_m(n)$	$r_m(n)$	mth estimate of aperiodic signal
$D_m(k)$	$R_m(k)$	mth estimate of aperiodic spectrum
$g_m(n)$	$p_m(n)$	mth estimate of periodic signal (p. 2)
$g_m(n)$	$g_m(n)$	mth estimate of periodic signal (p. 5)
$\delta(n)$	$l(n)$	hypothetical signal
$\Delta(k)$	$L(k)$	hypothetical spectrum
$\gamma(n)$	$h(n)$	hypothetical periodic signal
$?(k)$	$H(k)$	hypothetical periodic spectrum

Table D.1: Key to symbols used here and by YAD (Yegnanarayana et al. 1998).

Article	Description of results
d'Alessandro et al. (1995)	Comparison of decomposed aperiodic component with original noise component of one synthetic signal.
Darsinos et al. (1995)	Comparison of prescribed and measured HNR as a function of jitter (0–5 %) and shimmer (0–1.5 dB) on a synthetic signal.
Richard and d'Alessandro (1997)	An example of modification of an aperiodic component, extracted from recorded speech. Three examples of decomposed sentences
Yegnanarayana et al. (1998)	Two examples of decomposed synthetic speech compared with its constituents: one with constant noise, one with modulated noise. Two spectrograms of decompositions of recorded phrases ('ce voyage ...' and 'Je pense que ...').
d'Alessandro et al. (1998)	An example of decomposed synthetic speech compared with its constituents (showing modulated noise). The effects of jitter, shimmer and glides on measured HNR. A study of prescribed and measured HNR over a range of fundamental frequencies (80–300 Hz).

Table D.2: Summary of published PAPD results.

The decompositions of recorded speech claim to offer the capability of a continuous transition from normal voiced speech to whispered speech. Their examples demonstrate this to reasonable effect, although obvious remnants of voicing can still be heard. The decomposed aperiodic signal seems to give a fair imitation of whispered speech, but for generating breathy voice quality, this model is too simplistic. However, a realistic model would have to take into account the modified behaviour of the vocal folds. The modifications they made using the decomposed speech tended to introduce further audible artefacts.

D.2 Précis

The speech model assumes that an excitation signal $a(n)$, which is a superposition of periodic and aperiodic components, $g(n)$ and $d(n)$ respectively, is filtered by the vocal-tract with impulse response $q(n)$:

$$\begin{aligned} s(n) &= a(n) * q(n) \\ &= (g + d) * q\,, \end{aligned} \tag{D.1}$$

where $*$ denotes convolution. In the frequency-domain, the spectra are multiplied:

$$\begin{aligned} S(k) &= A(k)\,Q(k) \\ &= (G + D)\,Q\,. \end{aligned} \tag{D.2}$$

Figure D.1 is a schematic summary of the PAPD, which illustrates the way its algorithm is encased by an LPC analysis/synthesis shell (indicated by the dotted rectangle). This shell prewhitens the input signals before decomposition, and returns the spectral colouring (e.g., from the formants) afterwards. The algorithm operates on the excitation signal, $a(n)$, to separate the periodic (harmonic) and aperiodic (anharmonic) components in a two-stage process.

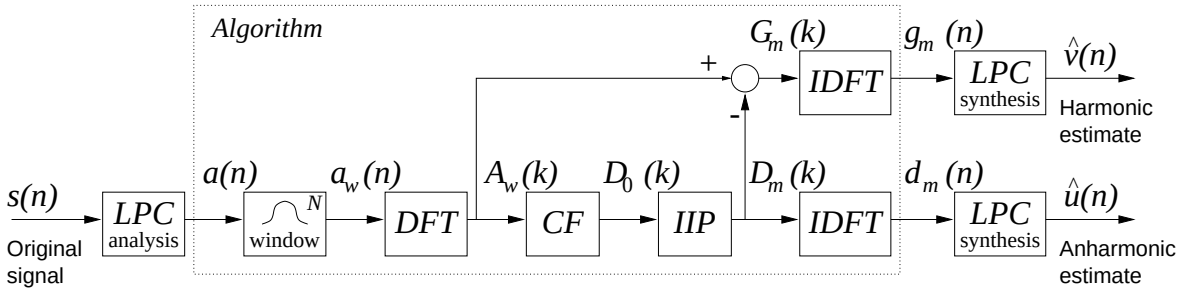


Figure D.1: The periodic-aperiodic decomposition (PAPD) algorithm, whose core comprises a cepstral filter (CF) and the iterative interpolation process (IIP).

The first stage makes an initial separation in the frequency domain using a cepstral filter. The signal $a(n)$ is windowed and zero-padded to form $a_w(n)$, where N is the DFT length and $(N/2 - 1)$ the window length. Its spectrum $A_w(k)$ and real cepstrum are computed (Noll 1967;

Deller et al. 1993, ch. 6). The real cepstrum is divided into three parts (see Yegnanarayana et al. 1998, Figs. 3 and 4): (i) the vocal-tract filter response (extracted by a low-time, rectangular lifter, 0–2 ms); (ii) the periodic excitation component (taken from the first rahmonic only, using a 1 ms fixed-width, band-pass, rectangular lifter);[1] and (iii) the aperiodic excitation component (the remainder). The periodic region of the cepstrum, part (ii), is extracted and its DFT is computed, to yield the log-spectrum. By comparing the periodic log-spectrum to zero, the bins of the spectrum $A_w(k)$ are assigned to either the periodic component (positive values) or the aperiodic component (negative values). The initial aperiodic estimate is thus set equal to the original spectrum for the aperiodic bins B_d and zero elsewhere:

$$D_0(k) = \begin{cases} A_w(k) & \text{for } k \in B_d, \\ 0 & \text{otherwise.} \end{cases} \tag{D.3}$$

The second stage is an iterative interpolation process (IIP), involving repeated transformations between FD and TD. The IDFT of $D_0(k)$ is not generally time-compact like $a_w(n)$: that is, $d_0 \neq 0$ for $N/2 \leq n \leq N$. The interpolation sets these points to zero, computes the DFT, resets $D_m(k) = A_w(k)$ for $k \in B_d$, computes the IDFT and so on. Setting the points to zero is equivalent to multiplying by a rectangular window: $\xi(n) = 1$ for $n \in \{1, 2, \ldots, N/2 - 1\}$, 0 for $n \in \{N/2, \ldots, N\}$. The process is repeated for 20 iterations, which Yegnanarayana et al. considered enough to allow $D_m(k)$ to converge.

D.2.1 Theoretical argument

Yegnanarayana et al. assume (Yegnanarayana et al. 1998, p.5) that "$R(k) = E(k)$ for $k \in F_r$" (col. 1, para 4; in our notation, $D(k) = C_w(k)$ for $k \in B_d$), which implies that the periodic spectrum $G(k)$ is precisely zero for those frequency bins. Using the argument of compactness that they employ in Eqs. 16 and 17 (col. 2, bottom), it can be seen that the spectrum is zero at all frequencies: $G(k) = 0 \; \forall k$. Yet, the authors remark that "the sidelobe effects of the windowing may produce significant values in the noise regions" (Yegnanarayana et al. 1998, p. 5). Therefore, provided that their argument is true, and that some part of the periodic component must reside in the aperiodic bins (as they remark), we would expect the convergent solution of the IIP to be the original spectrum: $D_m(k) \to A_w(k) \, \forall k$, as $m \to \infty$. In fact, the IIP, which is based on Parseval's theorem, is a standard signal reconstruction technique (Hayes et al. 1980).

However, Eqs. 12, 13 and 14 should not be strict inequalities, since $D_m(k)$ is equal to $D(k)$ at convergence.[2] So, while the expressions guarantee that the error does not increase, they alone cannot guarantee that they converge on a unique solution, a point noted in Hayes et al.

[1] In contrast to the method proposed by de Krom (1993).

[2] In the Papoulis-Gerchberg extrapolation technique from which this method is derived (Papoulis 1984), the convergence region is explicitly excluded from the proof for this very reason.

(1980). Let us suppose, therefore, that there are two possible outcomes of the IIP:

$$\lim_{m\to\infty} \quad D_m(k) = A_w(k)\,; \qquad \text{(Case 1)}$$

$$\lim_{m\to\infty} \quad D_m(k) \neq A_w(k)\,. \qquad \text{(Case 2)}$$

If Case 1 is true, we have a convergent solution, but our signal has not been decomposed into periodic and aperiodic components. If Case 2 is true, then we have a second solution $\Delta(k)$ that is time-compact and matches $D(k)$ for $k \in B_d$. Therefore, we could potentially have any linear combination of them, subject to $\Delta(k) = D(k) = A_w(k)$ for $k \in B_d$, and still meet these two criteria, i.e., $\mu A_w + (1-\mu)\Delta$ for any real value of μ. The particular solution (and hence the value of μ) would depend on precisely which bins $k \in B_d$ were chosen. The question then becomes, do our initial conditions $D_0(k)$ set a value of μ that yields the desired solution? Case 2 implies that the difference $?\,(k) = A_w - \Delta$ is non-zero, and thus, the convergence point would not be reached, no interpolation of $?$ would take place and the solution would be (somewhat arbitrarily) determined by the initial assignment of bins. It is not immediately obvious, in relation to $\gamma(n)$ being time-compact and $?\,(k)$ a comb-filtered spectrum, that "these two constraints cannot be satisfied simultaneously" (Yegnanarayana et al. 1998, p. 6). Time-compact signals with comb-like spectra exist, and two simple examples demonstrate this in Figure D.2. Nevertheless, if the two constraints cannot coexist, then $?\,(k) = 0$, $\mu = 0$, $\Delta(k) = A_w(k)\ \forall k$, which is Case 1 again.

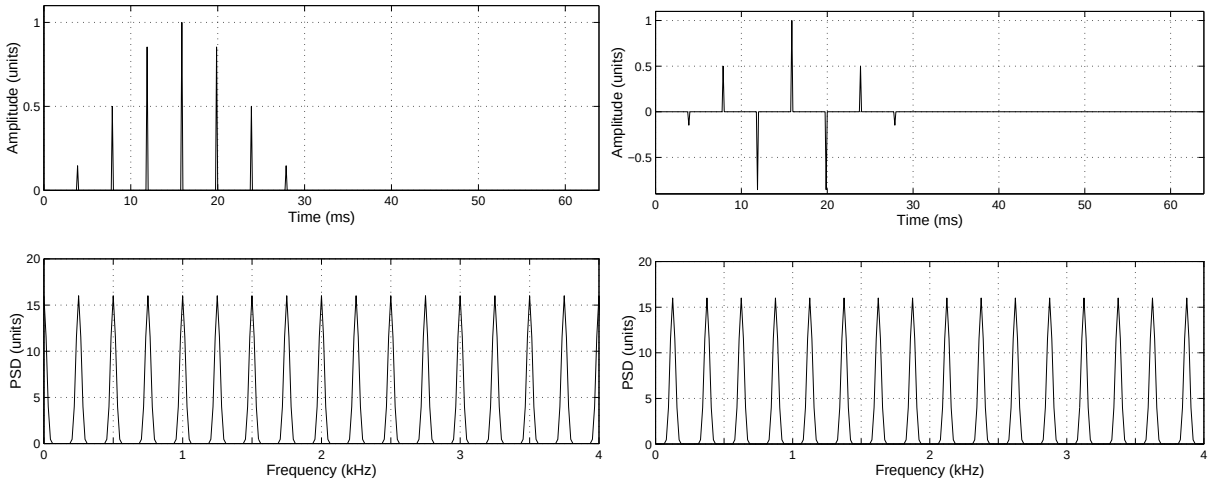


Figure D.2: Two examples of a time-compact signals (top) and their comb-like spectra (bottom). The signals are positive (left) and alternating (right) windowed pulse trains, and the resultant spectra are interleaved corresponding to the respective periodic and aperiodic bins of an excitation with $f_0 = 250\,\text{Hz}$.

D.3 Simulations

In their trials (Yegnanarayana et al. 1998; d'Alessandro et al. 1998), Yegnanarayana et al. used a Hamming window, $N = 512$ or 1024, and sampling rate $f_s = 8\,\text{kHz}$. The PAPD was evaluated by the measured HNR and a perceptual spectral distance. We ran simulations of the PAPD using their parameters ($N = 512$) on a mid-vowel section of the first [ɑ] of example 1 ($^{\#}1$) from Section 6.2.1, which was 6:1 downsampled to allow direct comparison with Yegnanarayana et al. (1998). The signal was LPC pre-emphasised (10 pole, autocorrelation) and 255 points were used for the analysis. At each iteration m of the interpolation, the signal power in the periodic and aperiodic estimates, $\langle d_m^2 \rangle$ and $\langle g_m^2 \rangle$, were calculated and plotted.

The results showed that the aperiodic estimate d_m began to approach convergence after about 1000 iterations, rather than after 20 as proposed (Yegnanarayana et al. 1998). Moreover, the solution upon which it appeared to converge was the original excitation signal, $a_w(n)$, suggesting that the algorithm, rather than decomposing the speech into periodic and aperiodic parts, actually reconstructed the original signal, using a subset of the Fourier coefficients (roughly one half). Repeating the tests at other parts of the utterance revealed the same behaviour. A second series of simulations was performed with signals synthesised from a pulse train plus Gaussian white noise (GWN) at HNRs ranging from -20 to $\infty\,\text{dB}$. Being spectrally flat, these signals required no LPC processing. Although convergence appeared to need a greater number of iterations, the results were similar: the IIP reconstructed the original signal, rather than effecting a stable signal decomposition.

Figure D.3 shows the effect of IIP on the decomposed components (top) and the PAPD performance (bottom), for a pulse train in GWN. Again, the parameters used were as specified in Yegnanarayana et al. (1998): 255-point Hamming window, 512-point DFT, and 8 kHz sampling rate. As with the other examples, the aperiodic estimate converged to the original signal, the periodic estimate to zero and the error to the original periodic component. The performance, despite showing a marginal improvement initially in this case, suffered severe degradation as the interpolation process was iterated, falling by 4 dB (η_v from 0 dB to $-4\,\text{dB}$, η_u from 5 dB to 1 dB). By comparison, the PSHF achieved performances of $\eta_v = 1\,\text{dB}$ and $\eta_u = 7\,\text{dB}$ on the same example. The effective reconstruction of the original signal from the initial aperiodic estimate $d_0(n)$ was consistently observed for all trials over a wide range of noise levels, with different f_0 values, DFT sizes and window functions. The initial conditions and the rate of convergence varied depending on the original signal's real cepstrum, which was governed by the choice of window and the details of the noise, but the asymptotic behaviour appeared in every case. Thus, because of the theoretical aspects that were overlooked, and the low number of iterations used, the PAPD algorithm appears to yield a reasonable decomposition.

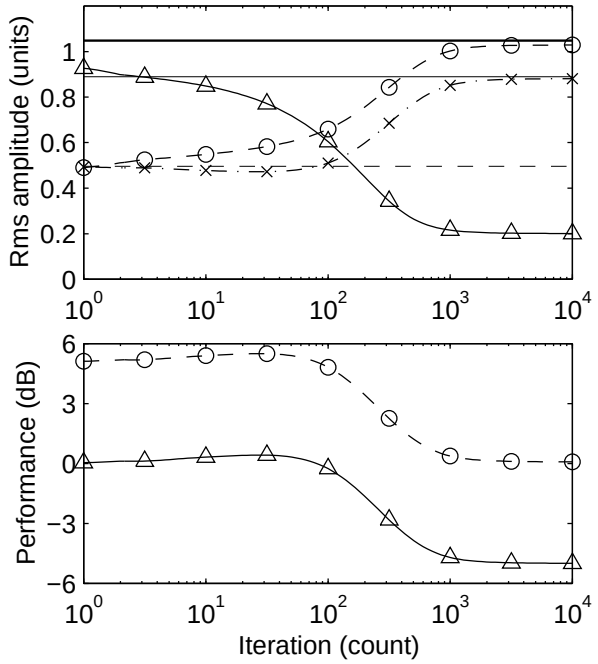


Figure D.3: Effect of the PAPD's iterative process. Top: log-linear plot of the root mean square amplitude of the periodic estimate g_m (△, solid), the aperiodic estimate d_m (○, dashed) and the error (×, dash-dot) versus iteration count for a pulse train ($f_0 = 120\,\mathrm{Hz}$) in Gaussian white noise (HNR = 5 dB). The horizontal lines indicate the original signal (thick, solid) with its components (thin): periodic (solid) and aperiodic (dashed). Bottom: the periodic (△, solid) and aperiodic (○, dashed) performance in dB.

D.4 Discussion

Our simulations, running over 10^5 iterations, support our theoretical argument that convergence is achieved when the original signal has been reconstructed (Section D.2.1). This result was expected since similar algorithms have successfully been applied to incomplete spectra as a solution to signal reconstruction problems (Hayes et al. 1980; Anderson 1994; Taratorin and Sideman 1994). The iterative algorithm is, however, a cumbersome, computationally-intensive method of signal reconstruction: for this purpose it could probably be replaced by a single-step calculation, like other such methods (Hayes et al. 1980) (involving the Hilbert transform or convolution with the complex spectrum of the rectangular window $\xi(n)$ that enforces the time-compactness criterion).

The initial estimate of the aperiodic part is based on the assumption that $D(k) = A_w(k)$ for $k \in B_d$ (Eq. D.10), but this does not take account of the effects of windowing, despite the authors' earlier remark confirming that it is an important issue. The PAPD crucially depends on the interplay between the spectral leakage of the rectangular window and the original Hamming window to determine the rate of convergence. Therefore, the amount of energy in the interpolated bins ($k \notin B_d$) depends on the number of iterations, on the way the time-compactness criterion is enforced (i.e., by rectangular window), on the HNR (owing to side-lobe leakage of the periodic part into the initial estimate $D_0(k) = A_w(k)$, for $k \in B_d$), and on the details of the aperiodic spectrum in the initial estimate. It does not depend on the amount of aperiodic energy any more than on the periodic energy. It is probably chance that the interpolated aperiodic energy appoximates the expected HNR values. Indeed, the discrepancies observed in Fig. 3 of d'Alessandro et al. (1998, Section IIIA, p.18) can be explained by this and by the decision to iterate twenty times.

Pre-whitening of the speech signal, which could be applied to any decomposition technique, is generally a good idea for sources with similar spectral tilts. It is inspired by proofs of optimality, which are given for sinusoids in GWN (Bretthorst 1988). In practice, despite increasing the computational load, it is not likely to add much benefit to the procedure, although it will flatten spectral humps from the formants. Let us consider the case of a voiced fricative, e.g., [z], where the voiced part has a strongly negative spectral tilt and, for the region up to about 8 kHz, the slope of the frication spectrum is broadly positive. The LPC inverse filtering will make the slope of the frication noise spectrum even more positive.

In frequency bands where there is a low HNR, voicing makes a negligible contribution and yet the PAPD allocates, on average, half of the excitation energy to the initial periodic estimate. Although the distinction between harmonics and noise is unclear, the PSHF allocates only one quarter of the excitation energy to the harmonic estimate in such bands. The authors note "that the decomposition algorithm is able to separate aspiration noises and the periodic

noise in the voice source" (Yegnanarayana et al. 1998, p. 9). However, the low sampling rate ($f_s = 8\,\text{kHz}$) used in Yegnanarayana et al. (1998) means that much of the turbulence noise was missed. These factors hindered the PAPD's ability to find new or hidden features in the decomposed speech.

In light of the above mathematical argument and the simulation results, we conclude that the PAPD ultimately converges, if the objective is a decomposition, to the wrong solution. Nonetheless, while there are some critical flaws and several shortcomings in the PAPD algorithm (Yegnanarayana et al. 1998), their use of synthetic signals and choice of variables offer a comprehensive methodology for testing decomposition algorithms, which we have largely followed.

D.5 Original statement of proof

The following is an extract from Yegnanarayana et al. (1998), which proposes a proof for the convergence of the PAPD algorithm. It is included here for ease of reference. Note that their symbols have been replaced with ours:

> Let N be the number of points in the DFT computation. Then the number of data samples should be less than or equal to $N/2$. In our case, we assume $N/2 - 1$ data samples (with $N/2$ even). Let $d(n)$ and $D(k)$ represent the true aperiodic component and its DFT, respectively, that we are trying to reconstruct by the iterative algorithm. Note that $d(n) = 0$, for $n \geq N/2$ and $D(k) = A_w(k)$, for $k \in B_d$ (noise regions in the frequency domain), where $A_w(k)$ are the DFT coefficients of the analysis segment of the LP residual.
>
> The iterative algorithm for reconstruction of the aperiodic component is as follows:
>
> *First Iteration:* We form the initial estimate of the DFT samples of the aperiodic component as
>
> $$D_0(k) = \begin{cases} A_w(k), & \text{for } k \in B_d \text{ (noise regions)} \\ 0, & \text{otherwise} \end{cases} \tag{D.8}$$
>
> and compute its IDFT $d_0(n)$. Since we have started with a segment of $N/2-1$ data samples, and we are using a N-point DFT, The time samples beyond $N/2 - 1$ are set to zero. That is, form a signal
>
> $$d_0'(n) = \begin{cases} d_0(n), & \text{for } n < N/2 \\ 0, & \text{otherwise.} \end{cases} \tag{D.9}$$

mth Iteration: Starting with $m = 1$, we compute the DFT $D'_{m-1}(k)$ of $d'_{m-1}(n)$, and form the function

$$D_m(k) = \begin{cases} A_w(k), & \text{for } k \in B_d \\ D'_{m-1}(k), & \text{otherwise} \end{cases} \tag{D.10}$$

and compute its IDFT $[d_m(n)]$. The time samples beyond $N/2 - 1$ are set to zero. That is

$$d'_m(n) = \begin{cases} d_m(n), & \text{for } n < N/2 \\ 0, & \text{otherwise.} \end{cases} \tag{D.11}$$

We shall now show that the functions $D_m(k)$ [tend] to $D(k)$ as $k \to \infty$, in the mean square (MS) sense.

Since $d'_m(n) = d_m(n)$, for $n < N/2$, and $d(n) = d'_m(n) = 0$, for $n \geq N/2$, we have

$$\sum_{n=1}^{N} |d(n) - d_m(n)|^2 > \sum_{n=1}^{N} |d(n) - d'_m(n)|^2 . \tag{D.12}$$

Likewise, in the frequency domain, since $D_{m+1}(k) = D(k) = A_w(k)$, for $k \in B_d$ (noise regions), and $D_{m+1}(k) = D'_m(k)$, for $k \notin B_d$, we have

$$\frac{1}{N} \sum_{k=1}^{N} |D(k) - D'_m(k)|^2 > \frac{1}{N} \sum_{k=1}^{N} |D(k) - D_{m+1}(k)|^2 . \tag{D.13}$$

Using Parseval's formula for the discrete case, we have the following result:

$$\begin{aligned} \frac{1}{N} \sum_{k=1}^{N} |D(k) - D_m(k)|^2 &= \sum_{n=1}^{N} |d(n) - d_m(n)|^2 \\ > \sum_{n=1}^{N} |d(n) - d'_m(n)|^2 &= \frac{1}{N} \sum_{k=1}^{N} |D(k) - D'_m(k)|^2 \\ &> \frac{1}{N} \sum_{k=1}^{N} |D(k) - D_{m+1}(k)|^2 . \end{aligned} \tag{D.14}$$

This result shows that excessive iterations will reduce the mean-square value of the error

$$\frac{1}{N} \sum_{k=1}^{N} |D(k) - D_m(k)|^2 . \tag{D.15}$$

Furthermore, the limit of this error is zero (i.e., the functions $D_m(k)$ [tend] to $D(k)$ as $m \to \infty$). The limit exists because the error is non-negative and decreasing.

Suppose that the limit is strictly positive (and not zero). In this case there exists a function $\Delta(k) \neq D(k)$, and the functions $D_m(k)$ tend to $\Delta(k)$ as $[m] \to \infty$. We have

$$\Delta(k) = \begin{cases} A_w(k) = D_m(k) = D(k), & \text{for } k \in B_d \\ \neq D(k), & \text{otherwise.} \end{cases} \tag{D.16}$$

In the time domain, we have also $\delta(n) = 0$ for $n \geq N/2$. This is because the functions $d'_m(n)$ tend to $\delta(n)$ as $[m] \to \infty$. Form the difference γ between δ and $[d]$. This function must satisfy

$$?\,(k) = \begin{cases} 0, & \text{for } k \in B_d \\ \neq 0, & \text{otherwise} \end{cases} \tag{D.17}$$

and $\gamma(n) = 0$ for $n \geq N/2$. In other words, γ must be a comb-filtered signal in the frequency domain, because it is zero for $k \in B_d$, and it must also be bounded in time to the interval $[1, N/2 - 1]$. Clearly, these two constraints can not be satisfied simultaneously. Therefore, $\Delta = D$, and the functions D_m converge to D.

Therefore, the iterations can be repeated until the difference between the energies of the aperiodic component $d'_m(n)$ for two successive iterations is below a prefixed threshold.

References

Glossary

$\#1$	example 1: [pɑzɑ] by PJ, from $\mathcal{C}3$
$\#2$	example 2: [ɑz:] by PJ, from $\mathcal{C}3$
$\#3$	example 3: [ɑz:] by PJ, from $\mathcal{C}5$
$\mathcal{C}$	speech corpus number, as described in Ch. 4
AR	auto-regressive
ARMA	auto-regressive moving-average
CEA	classical electrical analogue
CF	cepstral filter, Fig. D.1
DFT	discrete Fourier transform
dMRI	dynamic magnetic resonance imaging, see Ch. 3
EGG	electroglottograph, aka. laryngograph or Lx
EPG	electropalatograph
FD	frequency domain
GWN	Gaussian white noise
HF	harmonic filter, Fig. 5.1
HNR	harmonics-to-noise ratio (dB), defined in Section 4.3.1
IDFT	inverse discrete Fourier transform
IEE	Institution of Electrical Engineers, UK
IEEE	Institute of Electrical and Electronics Engineers, USA
IIP	iterative interpolation process, Fig. D.1
IIR	infinite impulse response (of a digital filter)
LMS	least mean-squares
LPC	linear predictive coding
ML	maximum likelihood
MSE	mean squared error
OQ	open quotient
PAPD	periodic-aperiodic decomposition, see Appendix D
PGG	photoglottograph

PI	power interpolation, Figs. 5.1 and 5.3
PSD	power spectral density (dB/Hz)
PSHF	pitch-scaled harmonic filter, see Ch. 5
RMS	root mean-square
RP	received pronunciation
SER	signal-to-error ratio (dB), defined in Section 5.5.2
SPL	sound pressure level
STFT	short-term Fourier transform
STP	short-term power, defined in Section 7.1.3
TD	time domain
TF	transfer function
WT	wavelet transform, Fig. 5.8
VOAC	vocal-tract acoustics program, see Ch. 2
VTTF	vocal-tract transfer function

Bibliography

Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh, UK: Edinburgh University Press.

Allen (1953). *Phonetics in Ancient India.* as cited by Dixit (1983).

Alwan, A., S. Narayanan, and K. Haker (1997). Toward articulatory-acoustic models for liquid approximants on MRI and EPG data. Part II. The rhotics. *Journal of the Acoustical Society of America 101*(2), 1078–1089.

Anderson, J. C. (1994). Complex signal reconstruction from time-frequency magnitude. *IEEE Transactions on Acoustics, Speech and Signal Processing 6*, 297–300.

Awan, S. N. and M. L. Frenkel (1994). Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice 8*(3), 255–262.

Badin, P. (1991). Fricative consonants: acoustic and X-ray measurements. *Journal of Phonetics 19*, 397–408.

Badin, P., D. Beautemps, R. Laboissière, and J.-L. Schwartz (1995). Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model. *Journal of Phonetics 23*, 221–229.

Baer, T., J. C. Gore, L. C. Gracco, and P. W. Nye (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America 90*(2), 799–828.

Barney, A. M., C. H. Shadle, and P. O. A. L. Davies (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory. *Journal of the Acoustical Society of America 105*(1), 444–455.

Beautemps, D., P. Badin, and R. Laboissière (1995). Deriving vocal-tract functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. *Speech Communication 16*, 27–47.

Bendat, J. S. and A. G. Piersol (1984). *Random Data: Analysis and Measurement Procedures.* New York, NY: Wiley-Interscience.

Beranek, L. L. (1954). *Acoustics* (1st ed.). New York, NY: McGraw-Hill.

Bickley, C. and K. N. Stevens (1986). Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies. *Journal of Phonetics* (14), 373–382.

Blomgren, M., Y. Chen, M. L. Ng, and H. R. Gilbert (1998). Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *Journal of the Acoustical Society of America 103*(5 Pt. I), 2649–2658.

Bretthorst, G. L. (1988). *Bayesian Spectrum Analysis and Parameter Estimation.* Berlin, Germany: Springer-Verlag.

Bristow, G. (Ed.) (1984). *Electronic Speech Synthesis.* London, UK: Granada.

Brown, J. C. and M. S. Puckette (1993). A high resolution fundamental frequency determination based on phase changes of the Fourier transform. *Journal of the Acoustical Society of America 94*(2, Pt. 1), 662–667.

Catford, J. C. (1977). *Fundamental Problems in Phonetics.* Edinburgh, UK: Edinburgh University Press.

Childers, D. G. (Ed.) (2000). *Speech Processing and Synthesis Toolboxes.* New York, NY: Wiley.

Childers, D. G., J. M. Naik, J. N. Larar, A. K. Krishnamurthy, and G. P. Moore (1983). Electroglottography, speech, and ultra-high speed cinematography. *Proceedings of the International Conference on Physiology and Biophysics of the Voice,* Iowa City, IA, 202–220.

Coker, C. H., M. H. Krane, B. Y. Reis, and R. A. Kubli (1996). Search for unexplored effects in speech production. *Proceedings of the International Conference on Spoken Language Processing 1996,* Philadelphia, PA *14*(6), 415–422.

Cook, P. (1991). Noise and aperiodicity in the glottal source: A study of singer voices. *Proceedings of the 12th International Congress on Phonetic Science,* Aix-en-Provence, France, 166–170.

Cranen, B. (1991). Simultaneous modelling of EGG, PGG, and glottal flow. *Vocal Fold Physiology,* eds. J. Gauffin and B. Hammarberg, Singular Publishers, San Diego, CA, 57–64.

Crow, S. C. and F. H. Champagne (1971). Orderly structure in jet turbulence. *Journal of Fluid Mechanics 48*, 547–591.

Curle, N. (1955). The influence of solid boundaries upon aerodynamic sound. *Proceedings of the Royal Society,* London *A231*(1887), 505–514.

d'Alessandro, C. R. (1990). Time-frequency speech transformation based on an elementary waveform representation. *Speech Communication 9*(5/6), 419–431.

d'Alessandro, C. R., V. Darsinos, and B. Yegnanarayana (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing 6*(1), 12–23.

d'Alessandro, C. R., B. Yegnanarayana, and V. Darsinos (1995). Decomposition of speech signals into deterministic and stochastic components. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Detroit, MI, 760–763.

Damper, R. I., J. R. Thorpe, and C. H. Shadle (1995). Separation of speech from simultaneous talkers. *Proceedings of the 13th International Congress on Phonetic Sciences,* Stockholm, Sweden *3*, 282–285.

Dang, J. and K. Honda (1997). Acoustic characteristics of the piriform fossa in models and humans. *Journal of the Acoustical Society of America 101*(1), 456–465.

Dang, J., C. H. Shadle, Y. Kawanishi, K. Honda, and H. Suzuki (1997). An experimental study of the open end correction coefficient for side branches within an acoustic tube. *Journal of the Acoustical Society of America.*

Darsinos, V., C. R. d'Alessandro, and B. Yegnanarayana (1995). Evaluation of a periodic/aperiodic speech decomposition algorithm. *Proceedings of Eurospeech 1995,* Madrid, Spain, 393–396.

Davies, P. O. A. L. (1988). Practical flow duct acoustics. *Journal of Sound and Vibration 124*(1), 91–115.

Davies, P. O. A. L. (1991). Program suite VOAC. (unpublished) Data sheets 1 and 2, Institute of Sound and Vibration Research.

Davies, P. O. A. L., J. B. Bento Coelho, and M. Bhattacharya (1980). Reflection coefficients for an unflanged pipe with flow. *Journal of Sound and Vibration 72*(4), 543–546.

Davies, P. O. A. L., R. S. McGowan, and C. H. Shadle (1993). Practical flow duct acoustics applied to the vocal tract. *Vocal Fold Physiology: Frontiers in Basic Science,* ed. I. R. Titze, Singular Publishers, San Diego, CA, 93–142.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research 36*, 254–266.

Dejonckere, P. H. and J. Lebacq (1996). Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. *Journal for Oto-rhino-larynngology and its related Specialties 58*(6), 326–332.

Deller, J. R., J. G. Proakis, and J. H. L. Hansen (1993). *Discrete-time Processing of Speech Signals.* New York, NY: Macmillan.

Dixit, R. P. (1983). On defining aspiriation. *Proceedings of the 13th International Conference of Linguisitics,* Tokyo, Japan, 606–610.

Donoho, D. L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. *Proceedings of the Symposium in Applied Mathematics 47*, 173–205.

Dworkin, J. P. and R. J. Meleca (1997). *Vocal Pathologies.* San Diego, CA: Singular Publishers.

Fant, G. (1960). *Acoustic Theory of Speech Production.* The Hague, Netherlands: Mouton.

Fant, G. (1973). *Speech Sounds and Features.* Cambridge, MA: MIT Press.

Feder, M. (1993). Parameter estimation and extraction of helicopter signals observed with a wide-band interference. *IEEE Transactions on Signal Processing 41*(1), 232–244.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception* (2nd ed.). Berlin, Germany: Springer-Verlag.

Flanagan, J. L. and L. Cherry (1969). Excitation of vocal-tract synthesizers. *Journal of the Acoustical Society of America 45*(3), 764–769.

Frazier, R. H., S. Samsam, L. D. Braida, and A. V. Oppenheim (1976). Enhancement of speech by adaptive filtering. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, 251–253.

Gabelman, B., J. Kreiman, B. R. Gerratt, N. Antonanzas-Barroso, and A. Alwan (1998). Perceptually-motivated modeling of noise in pathological voices. *Proceedings of the joint International Congress on Acoustics and Meeting of the Acoustical Society of America,* Seattle, WA *2*, 1293–1294.

Graf, J. T. and N. Hubing (1993). Dynamic time warping comb filter for the enhancement of speech degraded by white Gaussian noise. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Minneapolis, MN *2*, 339–342.

Griffin, D. W. and J. S. Lim (1984). A new pitch estimation algorithm. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* San Diego, CA *67*, 592–601.

Griffin, D. W. and J. S. Lim (1985). A new model-based speech analysis/synthesis system. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Tampa, FL *2*, 513–516.

Griffin, D. W. and J. S. Lim (1988). Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing 36*(8), 1223–1235.

Hanson, D. G., J. J. Jiang, J. Chen, and B. R. Pauloski (1997). Acoustic measurement of change in voice quality with treatment for chronic posterior laryngitis. *Annals of Otology, Rhinology and Laryngology 106*(4), 279–285.

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America 101*(1), 466–481.

Hardcastle, W. J. and J. Laver (Eds.) (1997). *The Handbook of Phonetic Sciences*. Oxford, UK: Blackwell.

Hardwick, J., C. D. Yoo, and J. S. Lim (1993). Speech enhancement using the dual excitation speech model. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Minneapolis, MN *2*, 367–370.

Hayes, M. H., J. S. Lim, and A. V. Oppenhiem (1980). Signal reconstruction from phase or magnitude. *IEEE Transactions on Acoustics, Speech and Signal Processing 28*(6), 672–680.

Haywood, R. W. (1968). *Thermodynamic Tables in SI (metric) Units* (2nd ed.). Cambridge, UK: Cambridge University Press.

Heinz, J. M. and K. N. Stevens (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America 33*(5), 589–596.

Heinz, J. M. and K. N. Stevens (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. *Proceedings of the International Congress on Acoustics 1*(A44).

Henke, W. L. (1966). *Dynamic Articulatory Model of Speech Production using Computer Simulation*. Ph. D. thesis, MIT, Cambridge, MA.

Hermes, D. J. (1988). Measurement of pitch by sub-harmonic summation. *Journal of the Acoustical Society of America 83*(1), 257–264.

Hermes, D. J. (1991). Synthesis of breathy vowels: some research methods. *Speech Communication 10*(5-6), 497–502.

Herzel, H. (1993). Bifurcations and chaos in voice signals. *Applied Mechanical Review 46*(7), 399–413.

Hess, W. (1983). *Pitch determination of speech signals: algorithms and devices*. Berlin, Germany: Springer-Verlag.

Hillenbrand, J. (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech and Hearing Research 30*, 448–461.

Hirose, H. and S. Niimi (1987). The relationship between glottal opening and the transglottal pressure differences during consonant production. *Laryngeal Function in Phonation and Respiration*, 381–390.

Holzrichter, J. F., G. C. Burnett, and L. C. Ng (1998). Speech articulator measurements using low power EM-wave sensors. *Journal of the Acoustical Society of America 103*(1), 622–625.

Horii, Y. (1979). Fundamental frequency perturbations observed in sustained phonation. *Journal of Speech and Hearing Research 22*, 5–19.

Howard, D. M. (1999). The human singing voice. *Proceedings of the Royal Institution of Great Britain 70*, 113–134.

Howard, D. M. and A. J. Fourcin (1983). Instantaneous voice period measurement for cochlear stimulation. *Electronics Letters 19*(19), 776–778.

Ishizaka, K. and J. L. Flanagan (1972). Synthesis of voiced sound from a two-mass model of the vocal chords. *Bell Systems Technical Journal 51*, 1233–1268.

Jackson, P. J. B. (1997). Defining, measuring and modelling aspiration in speech. Twelve Month Report PJBJ-970729-C, Department of Electronics and Computer Science, University of Southampton, UK.

Jackson, P. J. B. (1998). *Nephthys project*. Southampton, UK: Department of Electronics and Computer Science, University of Southampton. http://www.isis.ecs.soton.ac.uk/research/projects/nephthys/.

Jackson, P. J. B. and C. H. Shadle (1998). Pitch-synchronous decomposition of mixed-source speech signals. *Proceedings of the joint International Congress on Acoustics and Meeting of the Acoustical Society of America,* Seattle, WA *1*, 263–264.

Jackson, P. J. B. and C. H. Shadle (1999a). Analysis of mixed-source speech sounds: aspiration, voiced fricatives and breathiness. *Proceedings of the International Conference on Voice Physiology and Biomechanics,* Berlin, Germany, 30.

Jackson, P. J. B. and C. H. Shadle (1999c). Modelling vocal-tract acoustics validated by flow experiments (abstract). *Journal of the Acoustical Society of America 105*(2, Pt. 2), 1161.

Jackson, P. J. B. and C. H. Shadle (2000a). Aero-acoustic modelling of voiced and unvoiced fricatives based on mri data. *Proceedings of the 5th Speech Production Seminar,* Seeon, Germany, 185–188.

Jackson, P. J. B. and C. H. Shadle (2000b). Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *Journal of the Acoustical Society of America 108*(4), 1421–1434.

Jackson, P. J. B. and C. H. Shadle (2000c). Performance of the pitch-scaled harmonic filter and applications in speech analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Istanbul, Turkey *3*, 1311–1314.

Jackson, P. J. B. and C. H. Shadle (submitted April 1999b). Decomposing speech signals into their simultaneous voiced and unvoiced components. *IEEE Transactions on Speech and Audio Processing*.

Jenkins, G. M. and D. G. Watts (1968). *Spectral Analysis and its applications*. Time Series Analysis. San Francisco, CA: Holden-Day.

Jesus, L. M. T. and C. H. Shadle (2000). Characterizing spectral characteristics of European Portuguese fricatives. *Proceedings of the 5th Speech Production Seminar,* Seeon, Germany, 301–304.

Johnson, K. (1997). *Acoustic and Auditory Phonetics*. Blackwell.

Kent, R. D. and C. Read (1992). *The Acoustic Analysis of Speech*. San Diego, CA: Singular Publishers.

Kim, J. W. (1970). A theory of aspiration. as cited by Dixit (1983).

Kinsler, L. E., A. R. Frey, A. B. Coppens, and J. V. Sanders (1982). *Fundamentals of Acoustics* (3rd ed.). New York, NY: Wiley.

Kitzing, P. (1983). Simultaneous photo- and electroglottographic measurements of voice strain. *Proceedings of the International Conference on Physiology and Biophysics of the Voice,* Iowa City, IA, 221–229.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America 67*(3), 971–995.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*(3), 737–793.

Klatt, D. H. and L. C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America 87*(2), 820–857.

Koike, Y. (1969). Vowel amplitude modulations in patients with laryngeal diseases. *Journal of the Acoustical Society of America 45*, 839–844.

Ladefoged, P. (1985). *Computer Speech Processing*, Chapter 1. The Phonetic Basis for Computer Speech Processing. Prentice-Hall.

Ladefoged et al., P. (1976). The stops of Owerri Igbo. as cited by Dixit (1983).

Laroche, J., Y. Stylianou, and E. Moulines (1993). HNS: Speech modification based on a harmonic + noise model. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Minneapolis, MN *93*(2), 550–553.

Lieberman, P. (1961). Perturbations in vocal pitch. *Journal of the Acoustical Society of America 33*, 597–603.

Lighthill, M. J. (1952). On sound generated aerodynamically. I. General theory. In *Proceedings of the Royal Society,* London, Volume A211, pp. 564–587.

Lighthill, M. J. (1954). On sound generated aerodynamically. II. Turbulence as a source of sound. In *Proceedings of the Royal Society,* London, Volume A222, pp. 1–32.

Liljencrants, J. (1985). *Speech Synthesis with a Reflection-Type Line Analog.* Ph. D. thesis, KTH, Stockholm, Sweden.

Lim, J. S., A. V. Oppenheim, and L. D. Braida (1978). Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*(4), 354–358.

Lisker, L. and A. S. Abramson (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Acoustic Characteristics of Speech,* reprinted from *Word 20*(3), 527–565.

Löfqvist, A., L. L. Koenig, and R. S. McGowan (1995). Vocal tract aerodynamics in /aCa/ utterances: Measurements. *Speech Communication 16*, 49–66.

Logan, B. T. and A. J. Robinson (1997). Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Munich, Germany.

Mair, S. J. and C. H. Shadle (1996). The voiced/voiceless distinction in fricatives: EPG, acoustic and aerodynamic data. *Proceedings of the Institute of Acoustics 18*(9), 163–169.

Majid, F. (1997). *X Wavelet Packet Laboratory.* New Haven, CT: Department of Mathemeatics, Yale University. Version 1.3, http://www.math.yale.edu/pub/wavelets/software/xwpl/html/xwpl.html.

Markel, J. D. (1972). Digital inverse filtering - a new tool for formant trajectory estimation. *IEEE Transactions on Audio and Electroacoustics AU-20*, 129–137.

Masaki, S., M. Tiede, K. Honda, Y. Shimada, I. Fujimoto, and Y. Nakamura (1999). Dynamic MR imaging of laryngeal movement in production of voiced and voiceless stops. *Proceedings of the International Conference on Voice Physiology and Biomechanics,* Berlin, Germany.

McGowan, R. S. (1992). Tongue-tip trills and vocal-tract wall compliance. *Journal of the Acoustical Society of America 91*(5), 2903–2910.

McGowan, R. S., L. L. Koenig, and A. Löfqvist (1995). Vocal tract aerodynamics in /aCa/ utterances: Simulations. *Speech Communication 16*, 67–88.

Mermelstein, P. (1967). On the piriform recesses and their acoustic effects. *Folia Phoniatrica 19*, 388–389.

Mermelstein, P. (1971). Calculation of the vocal-tract transfer function for speech synthesis applications. *Proceedings of the 7th International Congress on Acoustics C13*(23), 173–176.

Michaelis, D., T. Gramss, and H. W. Strube (1995). Glottal-to-noise excitation ratio — a new measure for describing pathological voices. *Acta Acustica 81*, 700–706.

Mohammad, M. A. S. (1997). *Jaleel project*. Southampton, UK: Department of Electronics and Computer Science, University of Southampton. http://www.isis.ecs.soton.ac.uk/research/projects/jaleel/.

Mohammad, M. A. S. (1999). *Dynamic measurements of speech articulators using Magnetic Resonance Imaging*. Ph. D. thesis, Department of Electronics and Computer Science, University of Southampton, UK.

Morse, P. M. (1981). *Vibration and Sound*. New York, NY: Acoustical Society of America.

Morse, P. M. and K. U. Ingard (1968). *Theoretical Acoustics*. New York, NY: McGraw-Hill.

Motoki, K., P. Badin, X. Pelorson, and H. Matsuzaki (2000). A modal parametric method for computing acoustic characteristics of three-dimensional vocal tract model. *Proceedings of the 5th Speech Production Seminar,* Seeon, Germany.

Munjal, M. L. (1987). *Acoustics of Ducts and Mufflers*. New York, NY: John Wiley and Sons.

Murphy, P. J. (1999). Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *Journal of the Acoustical Society of America 105*(5), 2866–2881.

Murry, T., J. Large, and J. Dalgaard (1979). Vocal jitter in sung and spoken vowels. *Proceedings of the Acoustical Society of America BB3*, 29–44.

Muta, H., T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukuda (1988). A pitch-synchronous analysis of hoarseness in running speech. *Journal of the Acoustical Society of America 84*(4), 1292–1301.

Narayanan, S. and A. Alwan (1996). Parametric hybrid source models for voiced and voiceless fricative consonants. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Atlanta, GA *1*, 377–380.

Narayanan, S., A. Alwan, and K. Haker (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America 98*(3), 1325–1347.

Narayanan, S. S. (1995). *Fricative Consonants: An articulatory, acoustic and systems study.* Ph. D. thesis, Department of Electrical Engineering, University of California, Los Angeles, CA.

Noll, A. M. (1967). Cepstrum pitch determination. *Journal of the Acoustical Society of America 41*, 293–309.

Papoulis, A. (1984). *Signal Analysis.* New York, NY: McGraw-Hill.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America 60*(4), 911–918.

Paul, D. B. (1979). A robust vocoder with pitch-adaptive spectral envelope estimation and an integrated maximum-likelihood pitch estimator. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Washington D.C., 64–68.

Pelorson, X., G. C. J. Hofmans, M. Ranucci, and R. C. M. Bosch (1997). On the fluid mechanics of bilabial plosives. *Speech Communication 22*, 155–172.

Pierce (1981). *Acoustics. An Introduction to its Physical Principles and Applications.* Mechanical Engineering. New York, NY: McGraw-Hill.

Pinson, E. N. (1963). Pitch-synchronous time-domain estimation of formant frequencies and bandwidths. *Journal of the Acoustical Society of America 35*(8), 1264–1273.

Priestley, M. B. (1981). *Spectral Analysis and Time Series.* Probability and Mathematical Statistics. London, UK: Academic Press.

Qi, Y. and R. E. Hillman (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *Journal of the Acoustical Society of America 102*(1), 537–543.

Qi, Y., R. E. Hillman, and C. Milstein (1999). The estimation of signal-to-noise ratio in continuous speech for disordered voices. *Journal of the Acoustical Society of America, Letters to the Editor 105*(4), 2532–2535.

Richard, G. and C. R. d'Alessandro (1997). Modification of the aperiodic components of speech signals for synthesis. *Progress in Speech Synthesis,* eds. J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, Springer-Verlag, Berlin, Germany, 41–56.

Rife, D. C. and R. R. Boorstyn (1974). Single-tone parameter estimation from discrete-time observations. *IEEE Transactions on Information Theory 20*(5), 591–598.

Rife, D. C. and R. R. Boorstyn (1976). Multiple tone parameter estimation from discrete-time observations. *Bell Systems Technical Journal*, 1389–1410.

Rothenberg, M. R. (1973). A new inverse filtering technique for deriving the glottal air flow waveform during voicing. *Journal of the Acoustical Society of America 53*(6), 1632–1645.

Rothenberg, M. R. (1981). Acoustic interaction between the glottal source and the vocal tract. *Vocal Fold Physiology,* eds. K. N. Stevens and M. Hirano, University of Tokyo Press, 305–328.

Rothenberg, M. R. (1983). Source-tract acoustic interaction in breathy voice. *Proceedings of the International Conference on Physiology and Biophysics of the Voice,* Iowa City, IA, 465–481.

Rothenberg, M. R. (1992). A multichannel electroglottograph. *Journal of Voice 6*(1), 36–43.

Scott, R. J. and S. E. Gerber (1972). Pitch-synchronous time-compression of speech. *Proceedings of the Conference for Speech Communications Processing*, 63–65.

Scully, C. (1990). Articulatory synthesis. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Kluwer Academic.

Scully, C., E. Castelli, E. Brearley, and M. Shirt (1992). Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives. *Journal of Phonetics 20*, 39–51.

Scully, C. and S. J. Mair (1995). Relationships between different descriptive frameworks for plosive features of voicing and aspiration. *Levels in Speech Communication: Relations and Interaction,* eds. J. Schoentgen, J. M. Ramlot, C. Sorin, H. Meloni and J. Mariani, Elsevier Science B. V., Holland, 51–62.

Serra, X. and J. Smith (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition. *Computer Music Journal 14*(4), 12–24.

Shadle, C. H. (1985). *The acoustics of fricative consonants.* Ph. D. thesis, RLE Technical Report 506, MIT, Cambridge, MA.

Shadle, C. H. (1990). Articulatory-acoustic relationships in fricative consonants. *Speech Production and Speech Modelling,* eds. W. J. Hardcastle and A. Marchal, Kluwer Academic, Netherlands, 187–209.

Shadle, C. H. (1991). The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics 19*(3-4), 409–424.

Shadle, C. H. (1995a). Modelling the noise source in voiced fricatives. *Proceedings of the International Congress on Acoustics,* Trondheim, Germany *3*, 145–148.

Shadle, C. H. (1995b). Modelling the noise source in voiced fricatives. Technical Report, 1995/6 Research Journal, Department of Electronics and Computer Science, University of Southampton, UK.

Shadle, C. H., P. Badin, and A. Moulinier (1991). Towards the spectral characteristics of fricative consonants. *Proceedings of the 12th International Congress on Phonetic Sciences,* Aix-en-Provence, France *3*, 42–45.

Shadle, C. H., A. M. Barney, and P. O. A. L. Davies (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. II. Implications for speech production studies. *Journal of the Acoustical Society of America 105*(1), 456–466.

Shadle, C. H., C. U. Dobelke, and C. Scully (1992). Spectral analysis of fricatives in vowel context. *Journal de Physique IV Colloque C1, supplément au Journal de Physique III, Vol.2*(C1), 295–298.

Shadle, C. H. and S. J. Mair (1996). Quantifying spectral characteristics of fricatives. *Proceedings of the International Conference on Spoken Language Processing 1996,* Philadelphia, PA, 1517–1520.

Shadle, C. H., M. A. S. Mohammad, J. N. Carter, and P. J. B. Jackson (1999). Multi-planar dynamic Magnetic Resonance Imaging: New tools for speech research. *Proceedings of the International Congress on Phonetic Sciences,* San Francisco, CA *1*, 623–626.

Shadle, C. H. and C. Scully (1995). An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *Journal of Phonetics 23*, 53–66.

Shields, V. C. (1970). Separation of added speech signals by digital comb filtering. SM thesis, Department of Engineering, MIT, Cambridge, MA.

Shirai, K. and S. Masaki (1983). An estimation of the production process for fricative consonants. *Speech Communication 2*(2-3), 111–114.

Silva, F. M. and L. B. Almeida (1990). Speech separation by means of stationary least-squares harmonic estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Alberquerque, NM *2*, 809–812.

Simcox, C. D. and R. F. Hoglund (1971). Acoustic interactions with turbulent jets. *Transactions of the American Society of Mechanical Engineers, Journal of Basic Engineering*, 42–46.

Sinder, D. J. (1999). *Speech synthesis using an aeroacoustic fricative model.* Ph. D. thesis, Rutgers University, New Brunswick, NJ.

Sinder, D. J., M. H. Krane, and J. L. Flanagan (1998). Synthesis of fricative sounds using an aeroacoustic noise generation model. *Proceedings of the joint International Congress on Acoustics and Meeting of the Acoustical Society of America,* Seattle, WA *1*, 249–250.

Skoglund, J. and W. B. Kleijn (2000). On time-frequency masking in voiced speech. *IEEE Transactions on Speech and Audio Processing 8*(4), 361–369.

Sondhi, M. M. and J. Schroeter (1987). A hybrid time-frequency domain articulatory speech synthesiser. *IEEE Transactions on Acoustics, Speech and Signal Processing 35*(7), 955–967.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America 50*(4, Part 2), 1180–1192.

Stevens, K. N. (1993). Models for the production and acoustics of stop consonants. *Speech Communication 13*, 367–375.

Stevens, K. N. (1998). *Acoustic Phonetics.* Cambridge, MA: MIT Press.

Stevens, K. N. and S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America 64*(5), 1358–1368.

Stevens, K. N., S. E. Blumstein, L. Glickson, M. Burton, and K. Kurowski (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America 91*(5), 2979–3000.

Stevens, K. N. and H. M. Hanson (1995). Classification of glottal vibration from acoustic measurements. *Vocal Fold Physiology,* eds. O. Fujimura and M. Hirano, Singular Publishers Co., 147–170.

Stone, M. (1991). Imaging the tongue and vocal tract. *British Journal of Disorders of Communications 26*(1), 11–23.

Story, B. H. and I. R. Titze (1998a). Parameterization of vocal tract area functions by empirical orthogonal modes. *Journal of Phonetics 26*, 223–260.

Story, B. H. and I. R. Titze (1998b). Vocal tract area functions for an adult female speaker based on volumetric imaging. *Journal of the Acoustical Society of America 104*(1), 471–487.

Strope, B. P. and A. A. Alwan (1998). Amplitude modulation cues for perceptual voicing distinctions in noise. *Proceedings of the joint International Congress on Acoustics and Meeting of the Acoustical Society of America,* Seattle, WA *1*, 209–210.

Stylianou, Y. (1995). *Harmonic plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification.* Ph. D. thesis, Signals Department, ENST-Telecom, Paris, France. ftp://ftp.research.att.com/dist/stylianou/thesis.ps.gz.

Stylianou, Y., J. Laroche, and E. Moulines (1995). High-quality speech modification based on a harmonic + noise model. *Proceedings of Eurospeech 1995,* Madrid, Spain, 451–454.

Sundberg, J. (1977). The acoustics of the singing voice. *Scientific American*, 82–91.

Taratorin, A. M. and S. Sideman (1994). Signal reconstruction from noisy-phase and magnitude data. *Applied Optics 33*(23), 5415–5425.

Titze, I. R. (1994). *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall.

Titze, I. R. and B. H. Story (1997). Acoustic interactions of the voice source with the lower vocal tract. *Journal of the Acoustical Society of America 101*(4), 2234–2243.

White, P. R. (1997). An introduction to deterministic modelling. *Proceedings of the IEE Colloqium (Digest),* Stevenage, UK *9*, 2/1–2/3.

Wise, J. D., J. R. Caprio, and T. W. Parks (1976). Maximum likelihood pitch estimation. *IEEE Transactions on Acoustics, Speech and Signal Processing 24*, 418–423.

Yegnanarayana, B. (1981). Design of ARMA digital filters by pole-zero decomposition. *IEEE Transactions on Acoustics, Speech and Signal Processing 29*(3), 433–439.

Yegnanarayana, B., C. R. d'Alessandro, and V. Darsinos (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech and Audio Processing 6*(1), 1–11.

Yegnanarayana, B. and R. N. J. Veldhuis (1998). Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing 6*(4), 313–327.

Yoo, C. D. and J. S. Lim (1995). Speech enhancement based on the generalised dual excitation model with adaptive analysis window. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Detroit, MI, 832–835.

Yumoto, E., W. Gould, and T. Baer (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America 71*(6), 1544–1550.