UNIVERSITY OF SOUTHAMPTON

A Comparative Study of Two Area Function Derivation Techniques for Fricative Synthesis

by

Khazaimatol Shima Subari

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

February 2006

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Khazaimatol Shima Subari

It is still unclear to speech scientists how the brain is able to distinguish fricative sounds from the many cues within the speech signal. This thesis makes an acoustic and aticulatory investigation into the problem, using fricative utterances and magnetic resonance (MR) images. The main objective was to determine the perceptually important features of the fricative sound, by comparing the acoustic spectra and synthetic sounds derived from two area-derivation methods known as the Mermelstein technique and the Blum transform. They are distinct from each other in terms of the derivation and the use of the longitudinal axis of the tract and treatment of the sublingual cavity and pyriform sinuses. Thus, for each fricative articulation, two sets of area functions representing the vocal tract airway are derived, two sets of predicted spectra and consequently, two synthetic sounds. The degree of match between the measured and predicted spectra is quantified, and the synthetic and natural sounds are used in a perceptual experiment. This unique approach enables comparison of two sets of data to a single template and speculation about the features of the acoustic spectra and the articulatory aspects, that are perceptually important to the listener.

The results suggest that the cues to a fricative's identity lie in the higher frequency region. For the sibilant fricatives, these are the relatively higher amplitude levels above 3.0 kHz, while the distinct peak in the vicinity of 2.8 kHz gives the sound a "pitch quality" that was significantly perceived by the listeners. For the nonsibilant fricatives, the cues within the frication period did not provide sufficient information as to the fricatives' identity even though coarticulatory effects due to vowel context persisted; nevertheless, rough matches in amplitude levels were adequate to produce a synthetic version which listeners decided were comparable to the natural sound. Accurate measurements of the region posterior to the constriction were not essential for synthesis, supported by the fact that pyriform sinuses do not play an important role in modelling. The sublingual cavity was observed to affect the amplitude of the significant peak in /ʃ/ but it is sufficient to include it as an increase in area. The Mermelstein technique is proposed to be most suitable for fricative synthesis: it is easier to implement and side branches do not need to be modelled realistically.

Acknowledgements

All praises to the Creator for making this possible. My heartfelt gratitude to my two supervisors Prof Bob Damper and Dr Anna Barney who guided me patiently throughout the thesis with the most helpful suggestions, positive criticism, encouragement and support. I feel honoured to have worked with them. I wish to thank my former supervisor, Dr Christine Shadle, who mentored me through the first part of the investigation and who continued to be of great help whenever I needed it. Also, I am grateful to the EPSRC for funding my project.

I would like to thank Dr Matthew Wright for his helpful suggestions and comments during the mini-VIVAs, to Dr Philip Jackson for helping me out with ACTRA, and Dr Oliver Blacklock who always helped me on fricative issues which I felt too embarrassed to ask anyone else. I am grateful to Gordon Ramsey for supplying me with numerous articles that were no longer in print. I would like to thank Tasanawan Soonklang, for providing me the questionnaire forms from her experiments, which made the task of setting up my experiment easier. I would like to thank all of my friends and colleagues in ISIS and ECS, who were the subjects of my listening experiments—even if the first did not make it into the thesis. Their participation made the work enjoyable.

Thank you to Chas Tomlin, Amanda Goodacre and Tracy Cantlie for being the subjects of my experiment as well as the best admin staff ever. Thank you to Prof Mark Nixon Prof Eric Rogers and Eileen Simon, for taking the time to participate in my experiment despite their busy schedules and to Dr Peter Myerscough for getting me a job at MAC when I went into nominal. A special thank you to Hidayah for being the greatest office-mate one could ever have, and my best friend Nadia, who accompanied me to Turkey and whose friendship I could not have done without.

Finally, my deepest appreciation and love to my beloved mother and father who continuously offered their encouragement, love and support in every possible way. My siblings, especially Shaza, who was constantly there for me during turbulent times, while Zahir, Hafizah, Wahidah and Adilah can always be counted on to keep me in good spirits. A warm thank you for my grandmother whose prayers continuously encouraged me. Last but not least, a big hug for Armand, who is my strength and little source of inspiration.

To my mother,
my father,
my siblings,
and Armand,
with lots of love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This research seeks to make a contribution in the area of synthetic speech. To do so, it focuses on a group of English consonants known as fricatives. Their production requires the formation of a constriction in the oral region. Flow through the constriction generates turbulent noise, resulting in a sound that is stochastic in nature. Limited knowledge regarding the articulatory, acoustic and turbulence-generating mechanism as well as other aspects of the sound makes it difficult for speech scientists to determine the cues within the speech signal that are picked up and processed by the brain, allowing us to identify them easily during speech.

The objective is therefore to examine the features of the speech signal that are perceptually important to the listener. The approach to this problem is made from two different angles: articulatory configuration derived from magnetic resonance (MR) images and acoustic properties derived from the recorded speech of the same human subject. The main feature of this approach is the derivation of the vocal tract dimensions using two distinctively different methodologies, referred to as the Mermelstein technique and the Blum transform. The measurements of the vocal tract are presented as area functions, which are subsequently modelled as a concatenation of tubes by an acoustic modelling software named ACTRA to calculate the transfer functions. The combined product of the transfer function with a source model and a radiation characteristic term derives the predicted spectra and subsequently the synthetic sound of the fricative. The production of synthetic sounds using this technique is known as articulatory synthesis.

The Mermelstein- and Blum-derived spectra will be compared to the measured spectra of the natural utterances, and the differences between them will be quantified for analysis. Subsequently, the Mermelstein- and Blum-derived synthetic sounds will be compared to the naturally-derived sounds in a listening experiment conducted on 18 listeners. The results of the listening tests are analysed statistically for significance and, referred back to the fricative spectra and area functions to determine the important aspects of the sound. By taking this unique approach, we are, in effect, comparing two predicted spectra of the same fricative and two synthetically-derived sounds of the same fricative to a master template (i.e., the measured spectrum of the fricative and the naturally uttered fricative sound), which will allow us to speculate on the importance of the articulatory and acoustic aspects of the fricative sound, including the role of the sublingual cavity and pyriform sinuses.

The following sections provide the reader with background information on speech in general and fricatives in particular, information which is pertinent to this thesis. The motivations of this investigation, the objectives to be achieved and the questions to be answered are also identified. The final part of this chapter briefly explains the articulatory approach adopted and how each part of the process is laid out in the thesis.

1.1 Human Speech Production

The mechanism of speech production can be described as a coordination of three subsystems: the phonatory, the respiratory and the articulatory systems. Each system plays an essential role and contributes to final acoustic output.

The major organs of the respiratory system include the lungs and diaphragm, which control the air supply and thus affect sub-glottal pressure and rate of airflow in speech production. The phonatory system includes the trachea, larynx, pharyngeal cavity, nasal cavity and oral cavity. The articulatory system consists of fine anatomical features vital to speech production, such as the vocal folds, velum, jaw, teeth and lips. These components are collectively referred to as articulators. Coordinated movement of the articulators and simultaneous application of airflow from the lungs produce the desired speech sounds.

The main structure in which the propagating soundwave travels is known as the vocal tract. It begins at the larynx, extends through the pharynx, splits into the oral and nasal cavities and terminates at the lips and nose, respectively. It varies in length from person to person and has a variety of cross-sectional shapes that constantly change during speech by repositioning of the articulators. The cross-sectional areas along the length of the vocal tract can be represented by the "area function", i.e., area as a function of distance from the lips or the glottis.

Speech contains voiced or unvoiced components. The manner of acoustic excitation determines whether the sound is voiced or unvoiced. For voiced sounds, the larynx plays the important role of providing a periodic excitation known as a voice source. When airflow is forced through the glottis, the tensions in the vocal folds are adjusted so that they vibrate in an oscillatory fashion. The periodic interruption of sub-glottal airflow results in quasi-periodic puffs of air that excite the vocal tract. Examples of purely voiced sounds are vowels, which are characterised by well-defined resonances known as formants.

When a sound is classed as unvoiced, the vocal folds do not vibrate during production. Instead, the manner of excitation is created by the formation of a constriction somewhere along the vocal tract. The air is forced through the constriction, generating turbulence which acts as a noise source. Examples of sounds that come from such a source are whispering, aspiration, frication and plosives. For fricatives (i.e., frication sounds), a voice source as well as a noise source may be present during production. Therefore, there are two types of fricative consonants: voiced fricatives and unvoiced fricatives. This investigation will look exclusively at unvoiced fricatives.

1.2 World of Fricatives

In the English language, there are nine fricative phonemes. They are named after the articulator(s) that form the constriction and are given here in their phonetic representations. In the unvoiced case, there is a glottal fricative /h/ as in *hat*, a labiodental fricative /f/ as in *fan*, a dental fricative /θ/ as in *thick*, an alveolar fricative /s/ as in *son*, and a palato-alveolar fricative /ʃ/ as in *she*. The production of these fricatives creates a noise source downstream of the constriction and no voice source. The fricative /h/ does not have a voiced counterpart but when a voice source is present, the voiced counterparts—so called because the manner of articulation is similar—of the other fricatives are /v/ as in *van*, /ð/ as in *than*, /z/ as in *zoo*, and /ʒ/ as in *measure*, respectively.

Typically, when a speaker produces a fricative sound, the air expelled by the lungs encounters a constriction and the fluctuations in air pressure give rise to turbulent flow conditions (typically with Reynolds number greater than 2500, Davies et al. 1993), generating the noise source. However, depending on the vocal tract configuration, constrictions may also produce *stops* or *affricates*. For English and several other languages, the constriction for a fricative sound occurs in the anterior region of the oral cavity, but this does not apply to all cases. The glottal fricative /h/, for instance, does not require a constriction formation, and, as such, will not be studied here. Additionally, the presence of a voicing source causes interaction between the sources (Fant 1960; Flanagan 1972) and since we will not be looking at voiced source models, the voiced fricatives are excluded from this investigation.

The production mechanism for each fricative in steady state conditions is described as follows. For the labiodental fricative /f/, the upper incisors are placed against the lower lip to form a constriction, while for the dental fricative /θ/, the tongue is placed between the upper and lower incisors to form the constriction. For both fricatives, the airstream emerging from the constriction is directed towards the upper lips which act as a surface for the generation of noise. For the alveolar fricative /s/, the tongue tip is elevated and placed against the alveolar ridge to form a narrow constriction, while for the palato-alveolar fricative /ʃ/, the tongue blade is elevated and rests against the hard palate in such a way that a long and narrow channel is formed behind the point of constriction, the tip of the tongue also being near the alveolar ridge. For both these fricatives, the airstream emerging from the constriction is directed towards the lower incisors which act as an obstacle for the generation of air turbulence (Shadle 1985; Stevens 1998).

Fricatives are often referred to as sibilants or nonsibilants. According to Catford (1977) sibilant fricatives are those whose jet impinges on an obstacle. These are the fricatives /s, ʃ, z, ʒ/. On the other hand, nonsibilants are those whose jet emerges on an obstacle surface. These are the fricatives /f, v, θ, ð/. Chomsky and Halle (1968) defined another classification system which groups the fricatives /f, s, ʃ, v, z, ʒ/ as strident fricatives and /θ, ð/ as nonstrident fricatives. The latter terms are more widely used in the literature but not according to the above definition. 'Sibilant' and 'strident', it seems, are often used interchangeably. For example, Stevens (1985) and Narayanan et al. (1995) referred to /s, ʃ/ as strident fricatives and /f, θ/ as nonstrident fricatives. To follow the correct notation, we will refer to /s, ʃ/ as sibilant fricatives and /f, θ/ as nonsibilant fricatives. This categorisation is important; because of the acoustic similarities that are observed in each category the discussions throughout this thesis will be divided into these two categories. The nonsibilant fricatives are also referred to as front fricatives because the constriction is formed in the front-most region of the oral cavity (e.g., Shadle et al. 1996).

1.3 Side Branches of Vocal Tract

As well as the main vocal tract, the articulation of some fricatives results in side branches to the tract. There sometimes exists a vacant area between the tongue and the lower incisors known as a sublingual cavity. For instance, when the tongue tip is raised for the production of /s/, it is likely that a sublingual cavity exists. However, depending on the speaker, the /s/ sound can also be produced by placing the tongue tip at the base of the lower incisors which, in this case, will diminish the presence of a sublingual cavity (Shadle 1991). For /ʃ/ the sublingual cavity is consistently formed because the tongue body is elevated to form the constriction, while, for /θ/ it does not exist because the tongue must be placed between the teeth to form the constriction. For /f/, the presence of a sublingual cavity is speaker dependent and influenced by the vocalic context (Narayanan et al. 1995; Shadle et al. 1996).

Apart from the sublingual cavity, there are also the pyriform sinuses (also known as the piriform fossa) which are two cone-shaped columns that bifurcate off the laryngeal wall, just anterior to the glottis. Similar to the sublingual cavity, their significance in fricative production is relatively unknown, although for vowels, they have been observed to introduce a system zero between approximately 4.0 to 5.0 kHz (Dang and Honda 1997).

The importance of including the sublingual cavity and the pyriform sinuses in the fricative model, the effects they may have on the fricative spectra and consequently, the perceptual implications they will have for the listener have yet to be determined. These issues will be addressed in this investigation. Note that throughout this thesis, the pyriform sinuses and the sublingual cavity will be regularly referred to as side branches.

On a final note, fricatives belong in the same class as the affricate and stop consonants to form the *obstruents*. Obstruents have waveforms that are aperiodic and low in amplitude levels in comparison to other sounds of speech. Fricative's spectral characteristics show relatively undefined formants and the most energy in the higher frequencies (particularly the sibilant fricatives). This statement may be rather vague, but fricatives, like other classes of phonemes, also have pole and zero frequencies that depend primarily on the vocal tract area function and, therefore, a parametric representation in terms of fixed frequencies will fail when used on subjects of differing vocal tract lengths. The nonsibilant fricatives /f, v, θ, ð/ are harder to distinguish from each other because of their shorter duration, lower amplitudes and variable spectral shapes (Shadle et al. 1996). Some fricatives are highly influenced by the vocalic environment, shifting the formants and anti-resonances. This has been shown to occur for /f/ and /s/ (Shadle and Scully 1995) and will be discussed more in the following sections. Specifically for /f/, the location of the tongue tip has been known to vary as a consequence of the vocalic environment (Shadle et al. 1996, Stevens 1998).

1.4 Fricatives in Literature

Investigations pertaining to fricatives have taken numerous approaches and include among others, analysis of the acoustic spectra (e.g., spectral moments, overall sound pressure level etc.), derivation of articulatory parameters and/or aero-dynamic measurements and perceptual experiments. Most studies use a combination of data for comparison and analysis. In general, the main objectives of these studies have been the same: to determine the features that can be used to describe the fricative properties parametrically across subjects and/or determine the features which allow then to be distinguished from each other. The following sections present a review of fricative studies related in some way to this investigation. They concern:

1. the articulatory properties of fricatives;

2. the effects of vowel context on the fricative; and

3. listening experiments using fricative sounds.

The subject of articulatory properties constitutes the foundation of this investigation. The area functions, to which the whole Chapter 2 is dedicated, are derived from images containing articulatory information. The articulatory information for each fricative determines, among others, the location and length of the constriction and the existence of the sublingual cavity. This is important information needed to derive the predicted spectra. Studies referred under this subject heading include those which attempt to explain differences in the acoustic spectra by understanding the change in articulatory configuration.

The second subject is the effects of vowel context on the fricative. Here is explained why the corpora, both image and speech, used for this investigation were recorded in the vowel context /a/. Particular focus is on the nonsibilant /f/, because this fricative is much influenced by the vocalic environment. Thus, the image and speech data for /f/ were also recorded in vowel contexts /i/ and /u/. Chapter 3 defines a procedure which will quantify the effects of vowel context on each fricative. Note that the influence of vowel context will not be studied in great depth here, but part of the perceptual experiment will require the listener to identify the fricative sound and from the results, determine whether the effects of coarticulation from vowel contexts provide some cues that may have helped the listener identify the fricative.

The third and final subject is perceptual experiments utilising fricative sounds. This relates to the listening experiments done for this investigation, which required 18 subjects to listen to a combination of synthetic and naturally-uttered fricative segments for categorisation. Thus, a review of previous listening experiments utilising fricative sounds will also be made, since the results of these experiments may be important for purposes of comparison.

1.4.1 Articulatory Properties

The area function is a one-dimensional (1-D) representation of the cross-sectional area of the three-dimensional (3-D) vocal tract; therefore, some information is lost in the 3-D to 1-D transformation; the shape of each cross-section is not reflected in the function. For purposes of calculating the transfer functions, this is not of utmost importance, because the wavelength of sound is large compared to the axial dimensions of the vocal tract and, thus, to a first order approximation, unaffected by the cross-sectional shape of the tract (Stevens 1985). However, by understanding the locations of each articulator during production and the shape of the cross-sectional areas in some regions, the area functions can be better explained and checked for consistency. For the reader who is not familiar with magnetic resonance imaging (MRI), electropalatography (EPG) and ultrasound, more information on each imaging technique will be described in the early sections of Chapter 2, which cover the topic of the area function and its derivation in more detail.

A pioneer study on the articulatory-acoustic aspect of fricatives using MR images was conducted by Narayanan et al. (1995) in an attempt to derive area functions and provide a morphological explanation of fricative production. Four speakers were required to repeat sustained fricative segments in the neutral vowel /ə/. The following results were observed: the subjects consistently exhibited a sublingual cavity for /f/ but not for /θ/; the labiodental /f/ shows greatest variability between speakers in terms of tongue location, and the cross-sections of the vocal tract were not symmetrical, including the constriction areas. Additionally, the sibilant fricatives were observed to display a constriction area resembling a slit, because of the concave shape of the tongue as it braces against the palatal vault, while the nonsibilant fricatives had a constriction area that was elliptical in shape.

The results of Narayanan et al. (1995) complemented those of Stone and Lundberg (1996), which utilised ultrasound and EPG patterns to determine the shapes of the tongue's surface and the tongue-palate relationship during the production of some vowels and consonants. The results from this study indicated that /s, θ/ belonged to a category characterised by a "complete groove", while /ʃ/ exhibited features categorised as "front-raising". The former category indicates a channel-like feature because the lateral margins of the tongue were elevated with respect to the mid-sagittal plane while the latter category indicates that the anterior and midline portions of the tongue were elevated during production. The surface tongue shapes made during fricative production were similar to the ones made during vowel production. However, the EPG patterns indicated that during fricative production, the two sides of the tongue were in an upward position relative to the middle portion. Thus, although the tongue shapes for fricative production were indistinguishable from those made during vowel production, the tongue groove was elevated and the tongue braced against the hard palate to form the constriction for air turbulence.

Subsequently, Shadle et al. (1996) used MRI to determine if effects of vowel context were captured in the static images of two human speakers as they sustained fricative sounds in vowel contexts /a, i, u/. Data on the position of the tongue as well as the cross-sectional shape of the vocal tract during fricative production were extracted from the images. The results showed that the subjects had differences in the size of the sublingual cavity and sometimes had no cavity at all for the fricative /f/ depending on the vowel context. This implies that acoustic differences between the fricatives /f/ and /θ/ cannot be explained solely on the basis of the existence of the sublingual cavity, as previously suggested by Narayanan et al. (1995). Differences in the articulatory configuration of the fricative in different vowel contexts suggest that the effects of coarticulation persist in sustained sounds and support the use of MRI for future speech studies involving fricatives.

An investigation by Engwall and Badin (2000) focused on the effects of coarticulation observed in MRI data on Swedish fricatives, and derived conclusions that were similar to those of Shadle et al. (1996). The area functions of sustained fricatives uttered in different vowel contexts were compared in terms of the size of the constriction area. It was observed that /i/ produced the narrowest air passage (at some point in the vocal tract) and that fricatives uttered in this context have the smallest constriction when compared to other vowel contexts, suggesting that the degree of constriction of a particular fricative is affected by the vocalic environment. The fricative /f/ seems to be least affected, showing a consistent mid-sagittal distance measurement (at the constriction point) across the different vowel contexts. The area functions for fricatives uttered in the /u/ context are much different to those uttered in /a/ and /i/ contexts, and the authors suggest that this may be caused by greater differences in vocal tract configuration, occurring mostly in the oral cavity region.

1.4.2 Influence of Vocalic Environment

Two investigations into the effects of the vocalic environment on articulatory properties have already been referred to in the previous subsection. Here, two more investigations on the effects of vowel context, this time on the acoustic properties of fricatives, are described. The results of these studies motivated the investigation of fricative data embedded in vowel contexts for this study, particularly /f/. The final part of Chapter 3 attempts to quantify the effects of vowel context on each fricative in three different vowel contexts, /a, i, u/.

Shadle et al. (1995) investigated the fricatives /s, ʃ, x, ç/ uttered by an American speaker, a French speaker and two German speakers. In the first recording, the fricatives were preceded by the vowels /a, i/ and sustained for 3 s. In the second recording, the fricatives were inserted into nonsense words [pV_1FV_2] and repeated 10 times on a single breath, where V_1 and V_2 were either /a, i, u/, and F is one of the four fricatives indicated above. Data from the first set of recordings were analysed by placing 8 non-overlapping windows, each 20 ms in length in the steady-state portion of the sound and averaging each frequency component of the discrete Fourier transform (DFT). Data from the second set were analysed using a method known as "ensemble averaging" which places a 20 ms Hanning window at the same position in the steady state portion of the fricative spectra of all the uttered tokens and averaging each frequency component. The spectra of the sustained utterances taken from the German subjects for the fricatives /x, ç/ show more variation than the other two subjects. The spectral structure of short and sustained tokens of the same fricative shows consistency for the same vowel context in each subject, which indicates that there are no differences between short and sustained tokens. The /u/ context seemed to affect the spectra the most, by lowering the formant frequencies and bandwidths.

The effects of /u/ context were re-examined by Shadle and Scully (1995) on the fricative /s/, because of conflicting data from previous studies: Shadle et al. (1992) observed from ensemble averaging (the technique explained in the previous paragraph) that the /u/ context affects the spectrum of /s/, but Scully et al. (1992) observed from aerodynamic measurements (i.e., volume flowrate and intraoral pressure) that /s/ was relatively insensitive to vowel context. Re-examination of the data led to the following conclusions: the vocalic context does not have any influence on the articulatory configuration but some lip rounding extends into the fricative, creating a whistle configuration, given the right conditions, consequently introducing a second source mechanism which is reflected in the spectrum. The longer distance from the constriction to the impingement location as a result of this configuration may also be another explanation for the differences in spectra, as this might cause the jet stream to slow down and widen causing the particle velocity at the location of noise generation to be lower and affecting the noise source characteristics. The area of lip constriction was predicted to be smaller than that of vocal tract constriction, but this information has yet to be verified by taking into account various other measurements (i.e., constriction length, cross-sectional shape).

1.4.3 Listening Experiments

The final and major part of the current investigation (presented in Chapter 7) involves a perceptual experiment, where 18 listeners were required to identify a combination of natural and synthetic fricative sounds. Therefore, for purposes of comparison, a few listening experiments which have also utilised fricative sounds are reviewed.

Harris (1958) conducted a listening experiment from representations which combined the noise from one spoken fricative-vowel syllable with the voiced portion of another. For example, the fricative portion of /s/ was recombined with the vocalic portion /a/, which was originally uttered in the context of another fricative (i.e., /f/). Presumably, the most important part of the sound would determine which fricative a listener should hear. It was observed from the results that the important cues for the sibilant fricatives /s/ and /ʃ/ were located within the noise segment, but the identification of the nonsibilant fricatives /f/ and /θ/ were primarily based on cues in the vocalic segment of the sound. This conclusion was made as the listeners were observed to judge both /f/ and /θ/ as /f/ if the vocalic portion belonged to the segment uttered in the /f/ context, but as /θ/ when the vocalic portion was uttered in the context of another fricative. In other words, the judgement for /f/ was highly dependent on the vocalic portion of the sound.

Whalen (1991) performed a more complex listening experiment with presentations consisting of recombined fricative-vowel-fricative sequences to determine if the cues of /s/ and /ʃ/ can also be found in the transition region. In each experiment, the results indicated that the listeners would choose the fricative that was supported by the transition region, even if the duration of the frication noise was shorter. In conclusion to this study, it was suggested that the cues of /s, ʃ/ were not only found in the noise but also on the amplitude, duration of noise and the information in the formant transitions.

The results of both these studies strongly suggest the importance of the acoustic boundaries for the identification of the fricative, particularly for /f, θ/. Indeed, Stevens (1985) also observed that acoustic boundaries play a significant role in the distinction between the nonsibilant-sibilant (/θ/-/s/) and the anterior-nonanterior (/s/-/ʃ/) fricatives, in an investigation utilising synthetic segments of the fricative sound in /a/ context. In general, the synthetic segments were created in such a way that the amplitude of the turbulent noise (in the higher frequency region) was gradually reduced from one stimulus to the next in the series based on the amplitude of the peak in the same high frequency region of the neighbouring vowel. For the first test, where the listeners had to choose between /s/ and /θ/, the results indicate that if the peak level of noise (at the fifth formant) was higher than the peak level of the vowel (at the fifth formant), then the listeners would choose /s/ over /θ/ and if the level of noise was lower than the vowel, then they would choose /θ/. For the second test, where the listeners had to choose between /s/ and /ʃ/, the listeners would choose /ʃ/ if the peak level of noise (at the third formant) was higher than that of the peak level of the vowel (at the same formant) and as /s/ if the peak level was lower. Thus, it was suggested that listeners made their decisions as to the identity of the sound based on the relative amplitudes of the noise and adjacent vowel.

1.5 Motivation

The main motivation of this investigation was the importance—and lack—of accurate articulatory data for fricative configurations. The need for better articulatory information on fricative geometry has been much emphasised in previous studies (e.g., Scully et al. 1992, Shadle et al. 1995). Additionally, there is still much to know about fricatives, mainly the cues of the speech signal that are perceptually important for their identification. Exactly which cue is perceptually important to the listener remains unclear to speech scientists and yet the language user is able to process the multitude of cues of the speech signal with ease to identify quickly the fricative sounds during speech. Suggestions have been made on the importance of the acoustic boundaries (e.g., Stevens 1985, Whalen 1991) and vowel context (e.g., Harris 1958, Scully et al. 1992, Shadle et al. 1996) and consequently, much care has been taken to analyse fricative corpora embedded in a vocalic environment, including those used for this study. The major focus of this investigation, however, does not lie on the acoustic boundaries or the effects of vowel context, but part of this work will allow us to speculate as to whether the effects of coarticulation, found to be preserved in the sustained speech and image data (Shadle et al. 1996) are indeed helpful to the listener. The main objective of this investigation is to analyse the perceptually important aspects of the fricative steady-state, and this is done by comparing the predicted spectra and sounds derived using two different area-derivation techniques, with their natural counterparts.

In general, limited knowledge of fricative consonants hinders the progress of technological advancements related to speech, because of the considerable role they play in the spoken language. Examples of this are speech synthesisers, which require recognition by humans, and speech recognition systems, which need to recognise human speech, and these applications can be extended, and will be greatly beneficial to the field of medicine, education, telecommunications and multimedia. One example of a technology that will directly benefit from accurate representations of the vocal tract geometries is the articulatory synthesiser, which attempts to model human speech production by use of vocal tract parameters (e.g., the area function) and requires less memory capacity than concatenative synthesisers. Because articulatory synthesisers are dependent on parameters that are derived from human speech production geometries, high resolution 3-D images of the vocal tract during articulation will be greatly beneficial for the system. With the emergence of MRI, this data can now be derived. It is now up to speech researchers to draw the full potential of this facility, but this still depends on better fricative modelling. It may be that the fricative sound itself is not dependant on the accuracy of the articulatory properties because of its turbulent nature.

Thus, two area-derivation techniques are proposed to derive the model. The methodology for deriving the area function for each technique is very different in terms of how the measurements are taken and how the side branches are treated. Taking this approach effectively determines the best way to model fricative sounds, because making comparisons between predicted spectra and measured spectra, and synthetic sound and natural sound enables investigation of the perceptually important acoustic features and the important articulatory aspects of the fricatives.

1.6 Objectives

The main objectives of the thesis are presented in order of importance.

1. *Determination of the perceptually important features of the spectra:*
   Deriving the area functions using two significantly different techniques will predict spectra which have been found to differ from one another even though they represent the same fricative sound (Subari et al. 2004). These differences, however small, may be picked up by listeners when the synthetic sounds are compared against a naturally uttered version and consequently influence their identification and rating.

2. *Determination of the importance of the side branches:*
   The Blum technique allows a systematic method for the derivation of the side branches of the vocal tract allowing these areas to be modelled realistically for the synthesis. How does inclusion of the side branches affect the predicted spectra? Consequently, what are the perceptual effects for the listener? There is a possibility that their inclusion may have little or no effect on the listener as a result of acoustic decoupling between the cavities on either side of the constriction or the smaller dimensions of the side branch in comparison to the overall dimensions of the vocal tract.

3. *Determination of the best area-derivation technique:*
   Which area-derivation technique is the best method to use for deriving the synthetic sound, the Mermelstein technique or the Blum transform? In essence, the best area-derivation technique is defined as the one whose acoustic predictions are most similar to the measured spectra, and which consequently produce synthetic sounds which most resemble that of the natural version. However, because of the nature of fricative production which involves jet turbulence, there is the possibility that fricative synthesis does not require accurate measurements of vocal tract geometry and that a good approximation of the areas is adequate to produce a realistic sounding fricative.

4. *Determination of the usefulness of MRI images for a "dynamic" study of the fricative:*
   Dynamics in this context is defined as the study of sustained fricative sounds embedded in a vowel context: Scully et al. (1992) observed that embedding fricatives in vowel contexts provide clues to formant transitions, which assist in the identification of the place of articulation, while Shadle et al. (1996) observed that these clues were preserved in the steady-state of sustained fricatives. Thus the dynamic aspect of this study is not based on vowel-consonant transitions but rather the fact that the speech and image corpora were recorded in a vocalic environment. Even though these effects are preserved, are static MR images adequate for a study of fricative dynamics? In other words, do the coarticulatory effects provide sufficient cues to the fricative's identity? These issues will be addressed towards the end of this investigation.

In the course of this investigation, several other aspects will also be covered which are subsidiary objectives.

1. *Determination of the effects of vowel context on the fricative spectra:*
   The image and speech corpora of /f/ were recorded in three different vowel contexts. The effects of coarticulation in the acoustic spectra, the area functions and the predicted spectra will be studied.

2. *Definition of a quantification procedure to measure the differences between two spectra:*
   Definition is given of a measurement procedure that utilises a combination of dynamic time warping and mean square error measurements and uses it to measure the differences in the acoustic spectra of fricatives in different vowel contexts. The results are compared with findings from the literature to check if they are in agreement. If they are, then the technique can be confidently applied to measure the spectral differences between the measured and predicted spectra of the same fricative.

The findings of the subsidiary objectives are presented and discussed as the investigation progresses, but the four major objectives listed above will be addressed towards the end of this thesis (i.e., in Chapter 8), when all the findings of the previous chapters are brought together and discussed to derive the final conclusions.

Through this investigation, we hope to make a significant contribution to the study of speech in general and fricatives in particular. This thesis is the first to compare area-derivation techniques for the synthesis of fricative sounds. This approach is quite unique: it might be said that the investigation is analogous to comparing the area functions and predicted spectra of two subjects to a master template and trying to determine the features of the spectra that can be used to describe the perceptually important features parametrically across subjects.

Additionally, this study is also the first to investigate the acoustic and perceptual effects of including side branches in a synthetic fricative model. While extensive work has been done with regards to area function derivation, almost all has ignored side cavities. Finally, it is the first work to define an effective and reliable technique able to quantify the effects of vowel context on the fricative spectra and the degree of match between predicted and measured spectra.

1.7 Scope of Thesis

A flowchart of the overall experimental procedure of this investigation is shown in Figure 1.1. In this figure, the description following each chapter title refers to the main contribution or topic of the chapter (not the chapter title). The details of each chapter will be described in more detail in the following section.

In general, this investigation is based on the acoustic analysis of a speech corpus which consists of vowel-fricative-vowel (hereafter denoted by [VFV]) utterances with the fricative segment sustained and the area derivation of an image corpus of the same subject, where the images captured with the fricative segment were also sustained. As previously stated, the area functions are derived using the Mermelstein technique and the

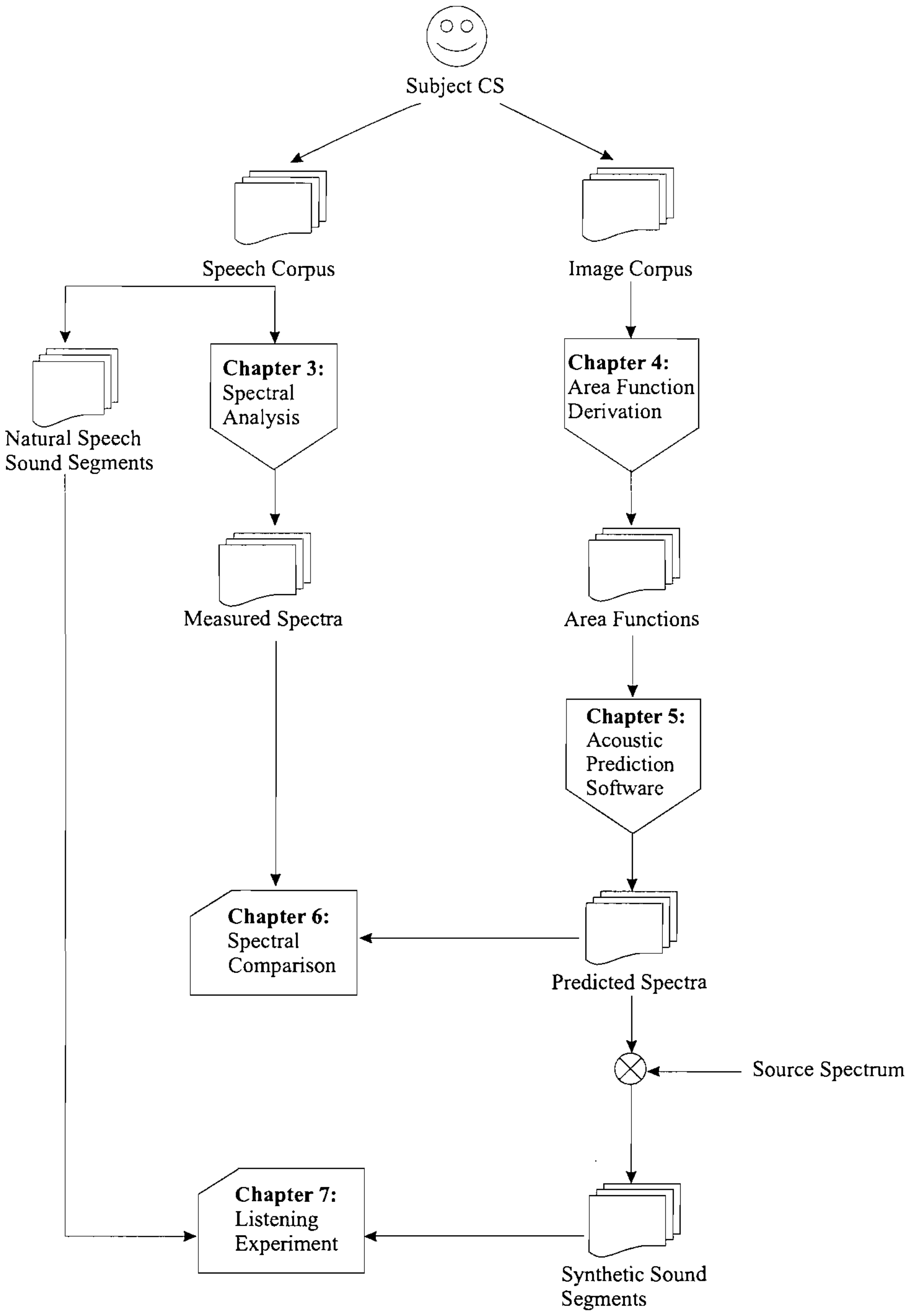
Subject CS
Speech Corpus
Image Corpus
Natural Speech Sound Segments
Chapter 3: Spectral Analysis
Chapter 4: Area Function Derivation
Measured Spectra
Area Functions
Chapter 5: Acoustic Prediction Software
Chapter 6: Spectral Comparison
Predicted Spectra
Source Spectrum
Chapter 7: Listening Experiment
Synthetic Sound Segments

FIGURE 1.1: Flowchart of overall experimental procedure. The speech and image corpora were recorded by a single subject, CS. The description following each chapter indicates the main topic/contribution of the chapter and not the chapter title itself.

Blum transform where the main differences between them are the derivation—and use—of the longitudinal axis of the vocal tract and the treatment of the side branches. The predicted spectra and synthetic sounds derived using the two area-derivation techniques are systematically compared by:

1. quantifying the differences between the measured and predicted spectra using a pre-defined quantification procedure, which utilises dynamic time warping (DTW) and mean square error measurements (MSE); and
2. analysing the results of a listening experiment which utilises a combination of synthetic sound segments derived from the predicted spectra and naturally recorded sound segments edited from the audio recordings.

The two points mentioned above are the major contributions of this investigation and are presented in Chapters 6 and 7 (see Figure 1.1).

In general, the articulatory and acoustic properties of voiceless fricative consonants, with some consideration of the effects of vowel context on the labiodental fricative /f/ are discussed. The effects of vowel context on *all* the voiceless fricatives will be measured in the final part of the acoustic analysis presented in Chapter 3, while the area functions of only /f/ in different vowel contexts will be derived. The differences in the area functions are the primary focus, as this will result in differences in the predicted spectra (Subari et al. 2004). Whether these differences are perceptually important will be determined by the results of the listening experiment, which, in short, requires the listener to match the 'best' synthetic segment to its natural counterpart. Statistical analyses by use of analysis of variance (ANOVA) and *t*-test will be applied to the results of the experiment to determine whether the decisions that were made by the listeners were significant.

In deriving the predicted spectra from the area functions, a distributed noise source is assumed (as opposed to a single concentrated source), to allow for a more realistic modelling of the fricative sound. However, the type of noise that will be used to represent the source will be somewhat simplified; the synthetic sounds will be derived by taking the combined product of the predicted spectra with the power spectrum of simple white noise. It was noted that this may not be an accurate representation of the frication source, as the noise turbulence that is involved in fricative sounds has been known to be a combination of types: monopole, dipole and/or quadrupole (Lighthill 1954; Narayanan and Alwan 2000).

1.8 Organisation of Thesis

This thesis is organised as follows. Chapter 2 presents topics relating to the image corpus, specifically the area function. The technology behind MRI, as well as the benefits and disadvantages of its use are compared with other data acquisition techniques. It explains why MRI was chosen to acquire the images for the investigation. Subsequently, studies based on derivation of the area function are presented in chronological order. It will give the reader a valuable insight into the development and the theoretical concepts of area-derivation techniques and, subsequently, introduce the reader to the two main area-derivation techniques implemented in the thesis.

Chapter 3 presents the acoustic analysis of the speech corpus derived using power spectral density (PSD). The articulatory-acoustic relationship is discussed as well as the variations within the fricative steady state. This chapter describes the quantification algorithm used to measure the degree of match between two spectra. Specifically for this part of the analysis, the algorithm allows quantification of the effects of vowel context on fricative spectra. The spectra derived from this part of the experiment will be referred to as the measured spectra.

Chapter 4 presents the analysis of the image corpus. It describes how the area-derivation techniques were implemented on the MR images and how the boundaries between the teeth and airway were determined. The extracted area functions are discussed and comparisons are made between the area functions derived from the two techniques and additionally, between the area functions of the labiodental /f/ in different vowel contexts. Subsequently, the area functions will be used as data for the acoustic prediction software known as ACTRA. The introduction to ACTRA and the computations that it is based on will be presented in Chapter 5.

Chapter 6 is the essence of the thesis and a merging point of all the results of the analyses that were made on the corpora. It describes how the predicted spectra are calculated and how the distributed source model is derived. The Mermelstein- and Blum-derived spectra will be compared with their measured counterparts and with each other. Further comparisons were made between spectra derived with and without inclusion of the areas of the side branches. These differences are further measured using the quantification procedure described in Chapter 3.

Chapter 7 presents the derivation of the synthetic sounds. It describes how the synthetic sound segments along with the natural sound segments of the audio recordings were combined to form the presentation of sounds used in the listening experiment conducted on the 18 subjects. ANOVA and *t*-test are used to determine whether the decisions made by the listeners are statistically significant. The main discussion is based on the differences between the sounds derived from the two area-derivation techniques and features of the spectra observed to be perceptually important to the listeners. The role of side branches in the modelling is also discussed. This enables a definition of the best area-derivation technique, leading to several important conclusions for the modelling of fricative sounds.

Finally, Chapter 8 summarises the work completed in this thesis, discusses the findings and derives the final conclusions from the results. The chapter ends with suggestions for future work.

1.9 Contributions

During the course of this investigation, the author made one poster presentation and published one article (listed below). Both are available for download at `http://www.ecs.soton.ac.uk/~kss01r/`.

- Subari, K. S. and C. H. Shadle (2004). An MRI and acoustic study of the effect of vowel context on fricatives. *UK Speech Young Researchers Meeting.* University College London, London, UK.
- Subari, K. S., C. H. Shadle, A. Barney and R. I. Damper (2004). Comparison of fricative vocal tract transfer functions derived using two different segmentation techniques. In *Proceedings of the International Conference on Signal Processing,* Istanbul, Turkey, pp. 164–168.

Chapter 2

Vocal Tract Measurements

This investigation is divided into two major parts: analysis of a speech corpus and analysis of an image corpus. The previous chapter introduced the reader to fricative speech sounds and their acoustic and articulatory properties, information that is related to the speech corpus. This chapter introduces the reader to magnetic resonance imaging (MRI), the area function and area-derivation techniques, information related to the image corpus. Comparisons are made between MRI and X-ray, electropalatography (EPG) and ultrasound, discussing the advantages and disadvantages of each data acquisition technique and consequently making it clear to the reader why MRI is specifically chosen to derive the image corpus. Additionally, the reader is made aware of the limitations of the technology, which may possibly affect the outcome of the experiment.

This is followed by a review of several area-derivation techniques in the literature and the reasons for selecting the Blum transform and Mermelstein technique. Since this chapter is primarily focused on vocal tract anatomy and area functions, a brief review is made on the role of pyriform sinuses in speech in general.

2.1 Data Acquisition Techniques for Speech Studies

Since its introduction for medical use in the early 1950's, MRI has proved to be very useful in speech research. This technology has enabled researchers to capture high resolution images of the vocal tract system during speech with good soft tissue resolution and separate images for each plane, which makes the images amenable to computerised 3-D modelling of the vocal tract (Narayanan et al. 1995; Holtrup 1998). Most importantly, the subject is not exposed to any known radiation risks.

Several recent speech studies have used MRI for speech research. A study of vowels was reported by Baer et al. (1991), followed by Moore (1992), Greenwood et al. (1992)

and Sulter et al. (1992); a study of nasals was done by Dang et al. (1993), a study of semi-vowels by Stone et al. (1996) and Zhang et al. (2003) and a study of fricatives by Narayanan et al. (1995), Shadle et al. (1996) and Holtrup (1998). Studies on area functions were made by Story et al. (1996) and Engwall and Badin (2000). Models in 3-D of the vocal tract shape have been constructed using MRI images by Tiede and Yehia (1997) and Badin et al. (1998). Additionally, a film on articulatory movements using MRI was reported by Foldvik et al. (1990).

The technology behind MRI is highly dependent on the hydrogen nucleus from which the signal originates. This nucleus, which consists of a single proton, has a net magnetisation of $\mathbf{M} = \mathrm{M}_x\mathbf{i} + \mathrm{M}_y\mathbf{j} + \mathrm{M}_z\mathbf{k} = 0$ in normal conditions where the spatial orientation of the moments are random. However, application of a static magnetisation field $\mathbf{B} = \mathrm{B}_o\mathbf{k}$—which, at 25°C is approximately 20,000 times the strength of the earth's magnetic field—forces the protons in the body to align in its direction, overcoming the thermal randomisation of the magnetic moments. Equilibrium is not reached instantly, but $\mathbf{M}$ grows slowly from 0 to $\mathrm{M}_o\mathbf{k}$. Once equilibrium is reached, a second, time-varying magnetic field, $\mathbf{B}_1 = \cos(\omega_o t)\mathbf{i} + \sin(\omega_o t)\mathbf{j}$, is applied in a plane transverse to $\mathrm{B}_o\mathbf{k}$. This field rotates at a radian frequency ω_o, causing the orientation of the magnetic dipoles to change direction and induces an electromotive force (EMF) in a receiver coil. The duration of $\mathbf{B}_1$ is chosen such that $\mathbf{M}$ is rotated into the transverse plane. The corresponding waveform is called the 90° excitation pulse with a frequency of $1/T_R$, where T_R is the repetition time. After application of the excitation pulse, the "relaxation time" for the dipoles to return to their original orientation is measured. The echo of the applied impulse is measured after the echo time, T_E.

In a single volume element of a magnetic resonance (MR) image, a single pixel consists of the net magnetisation of all dipoles at that region of the body. This usually involves a two step process. The first is application of the transverse magnetic field already described above. Apart from the hydrogen molecules, this process greatly influences the contrast of the tissues in the final image. Magnetisation not fully tilted at 90° may result in loss of signal; however it allows for shorter imaging times which are sometimes preferred (e.g., Greenwood et al. 1992). The second process encodes the spatial location of the signal during data acquisition. When the static field is a gradient field, the relaxation times can be related to the location. The proton density of a volume may then be calculated by use of the inverse Fourier transform (Holtrup 1998). In general there are two common imaging methods in MRI. The first is the "spin echo" (SE) technique, where a 90° excitation pulse is applied, followed by a 180° pulse. The second technique is known as the "turbo flash" (TF) technique where a single pulse is applied every T_R. For a more detailed explanation of the physics and functionality of MRI, refer to Wright (1997).

MRI does not come without disadvantages. There are in fact four major drawbacks which directly affect its use for speech studies:

1. MRI in general requires long acquisition times, which limits its use to sustained sounds. Good image resolution and thinner slices can be obtained using the SE technique at relatively long imaging times. It is possible to obtain images in a much shorter time (though still relatively longer compared to other imaging methods) using the TF technique. However, this is at the expense of poor resolution and thicker slices. Some studies suggest that long imaging times may cause fatigue in subjects. Consequently, the vocal tract may have altered in configuration during the imaging process. Depending on the application, either the TF or the SE technique may be more suitable.

2. The machine generates high ambient noise during operation and audio recordings made simultaneously with the scanning procedure are practically useless. Holtrup (1998) found it was impossible to extract good acoustic data from recordings made during an MRI session, particularly for the lower frequencies where the signal was completely submerged in noise. It is therefore more practical to make separate sound recordings when MRI is involved, although the vocal tract may not be in precisely the same articulatory configuration.

3. Subjects have to be in a supine position throughout the scanning procedure, a position which is not typical during normal speech. This may possibly alter the vocal tract configuration because of gravitational effects on the articulators (Tiede et al. 2000).

4. There is difficulty in distinguishing the boundaries between the airway and bone or teeth in MR images because of the low hydrogen content in these areas. Relying on a threshold value in image processing software to measure airway dimensions may incorrectly include space occupied by the teeth. Holtrup (1998) performed this step manually, bearing in mind that the computed airway must be anatomically possible. This solution to the air-teeth boundary problem is tenuous and a possible source of errors. Other researchers have used moulding materials (e.g., Narayanan et al. 1995) and superimposed teeth models on the vocal tract models. Discussion of this problem and the steps taken to overcome it in this investigation will continue in Chapter 4.

Before the 1990's, accessibility of MRI was limited, due to high operational costs and expensive equipment. Thus, the machinery was only available at selected hospitals. During this time it was common for researchers in speech studies to use X-ray radiography to obtain midsagittal profiles of subjects during speech. Apart from being readily accessible, relatively shorter imaging times allowed subjects to be filmed whilst saying full sentences. However, due to harmful biological effects of radiation, the subjects selected for such studies were usually already required to have X-ray sessions for medical reasons. This however, meant that subjects were not in a healthy condition during data acquisition and this could potentially have affected the results of the experiments.

Nowadays, the use of X-rays on normal subjects is viewed as unethical and dangerous, and there has been some concern that data recorded from (normal) speakers in the 1950-1970's may never be replicated. Following this, steps have been taken to ensure that the data is preserved (Munhall et al. 1994).

Electropalatography (EPG) is another method used in speech processing studies to acquire data while a subject is speaking. The EPG device consists of an acrylic palate which is custom-fitted to the subjects' palatal vault. Approximately 60-96 silver electrodes are embedded along the surface and the inner edges of the palate. Each time the tongue makes contact with the palate, signals are sent off and collected by a computer in real time. Speech therapists use this technique routinely to identify speaking disorders and help the patient correct articulation for specific sounds by mimicking the correct positioning of the tongue against the hard palate. In speech studies, the information it supplies is limited, because it does not provide detailed information of the vocal tract configuration or the tongue-tip position or shape. Therefore in articulatory-acoustic studies utilising EPG, another set of complementary data is used, such as ultrasound images (e.g., Stone and Lundberg 1996).

Ultrasound is an imaging tool that is widely used in the medical field. Ultrasound equipment is relatively inexpensive, portable, affordable and convenient for research laboratories and universities. Relatively detailed images are collected in real time and no radiation is involved in the process, making it a safe procedure. During an ultrasound scan, a transducer emitting high frequency acoustic waves is placed on the subject's body in the region of interest. The waves interact with the tissue and blood, reflected, and are picked up by the transducer which converts it back to electrical signals. The characteristic of the return signal (amplitude, phase, etc.) provides information on the nature of interaction and hence the type of medium in which they occurred. The speed of sound in tissue varies with tissue type, temperature and pressure and, assuming normal body temperature and pressure, the type of tissue provides information for the generation of images. This information is collected by a "beamformer" and processed for display (Quistgaard 1997).

However, ultrasound also has its disadvantages when it comes to speech studies. As the acoustic wave travels through the body, it is attenuated as parts of the signal are absorbed, scattered and/or reflected. Bone and air have the highest attenuation constants, therefore ultrasound images are not able to provide information on vocal tract shape during speech because there are limitations on where the transducer can be placed in the facial and neck region. For the same reason, not all information on tongue shape can be derived, as seen in the study by Stone and Lundberg (1996).

In contrast to MRI, X-ray and ultrasound have relatively shorter image acquisition times. Lateral X-ray and ultrasound procedures allow subjects to sit upright during scanning and both have relatively quieter machines, which allow for simultaneous recording of the speech sounds while the images are being taken. Hence it is not surprising that, although MRI does produce the best 3-D images with relatively high SNR, some researchers still prefer to investigate the vocal tract using these other methods as well.

2.2 Area Functions

The vocal tract resonances (or formant frequencies) which characterise the phonemes we perceive are dependent on the shape and dimension of the vocal tract. They are also affected by several other aspects, for example, coupling or decoupling between the cavities of the vocal tract (e.g., nasal cavity and oral cavity as in the production of nasals), differences in the compliance of the tract walls and radiation at the lips. In theory, acoustic modelling based on the area function of the vocal tract, measured entirely from the glottis up to the lips, is one method of estimating the formant frequencies. Indeed, several investigations have measured the area functions of the vocal tract, predicted the radiated acoustic spectra and subsequently compared them to spectra measured from real recordings (e.g., Mermelstein 1973; Goldstein 1980; Baer et al. 1991).

In the frequency range of interest to speech scientists, the wavelength of sound is large compared with the axial dimensions of the vocal tract. Therefore, to a first order approximation, only plane waves will propagate in the vocal tract. The wavefronts of the plane waves are normal to their direction of travel and tend to form equal angles with the walls of the vocal tract. To replicate the area encountered by the wavefront during wave propagation, the area measurements are read at regular intervals perpendicular to the longitudinal axis of the vocal tract. Each measured cross-sectional area is referred to as a grid plane (in a midsagittal diagram of the vocal tract, the grid planes are referred to as grid lines). The location and angle of the grid planes are determined by the area-derivation technique that is applied on the images.

2.2.1 Area-Derivation Techniques

Computation of the area function of the vocal tract has been carried out extensively by speech researchers. The techniques have been refined and redefined through the years, with advancements in technology and availability of new information. There are mainly two stages. Both are equally important and derivation may not necessarily be done in the same order. The first is deriving the longitudinal axis of the vocal tract. The second is the placement of the grid lines at regular intervals along the longitudinal axis. The grid lines, also known as "sagittal widths", are parameters commonly used to estimate the cross-sectional area using some kind of transformation. One well-established sagittal-to-area transformation is the $\alpha\beta$-model introduced by Heinz and Stevens (1965), and yet another is the recently published transformation proposed by Badin et al. (2005). For this investigation, the images are complete volume scans; by use of an image processing software known as 3D-Doctor, precise area measurements of the vocal tract's cross-sectional area can be obtained directly without sagittal-to-area transformations or any complementary data.

The review of this section will look specifically into the following aspects of each study:

1. the image acquisition technique;
2. the corpus;
3. the area-derivation technique;
4. the anatomical landmark(s) used to derive the area measurements; and
5. any supplementary data that were used to assist in the investigation.

Not all the issues listed above were addressed in each of the investigations stated below.

Coker and Fujimura (1966) introduced the concept of representing the vocal tract in terms of five variables and two circles, the larger circle representing the hard palate in a fixed location and the smaller one representing the tongue with a fixed radius and changing centre location. The variables represent the tongue position with respect to the larger circle, tongue tip and lower pharynx location, and opening and protrusion of the lips. Figure 2.1 shows a simplified version of this model. The paper does not address the issue of the derivation of the area function, but it provides the basis for further research in the field. Following this, the model was used to develop an articulatory speech synthesiser (Coker 1968; Coker 1976).

Heinz and Stevens (1965) introduced the $\alpha\beta$-model and by means of a simple analytical equation, can easily be adapted to suit different speakers. The model is given by $A(x) = \alpha d_j(x)^{\beta}$, where d_j is the sagittal-width of the jth slice and x is the distance of the grid plane from the glottis. To measure the vocal tract consistently across different subjects, three anatomical landmarks were used to determine the angles of the grid planes, illustrated in Figure 2.2. The anterior lower edge of the second cervical vertebra, denoted A, posterior nasal spine, B, and tip of crown of the most anterior maxillary central incisor, C, were connected in the centre, O, by three lines of equal distances. This centre point becomes the point of convergence for the radial slices. Following this, a straight line was drawn between B and C, and the midpoint, D, of this plane was found. Point D was then connected with a straight line to O and consequently, each segment within the planes AO and OD were segmented radially at 10° intervals while the remainder of the tract was sliced in parallel to AO and OD respectively at 10 mm intervals. Details of how the midline was computed were not given in the paper, but in instances where the midline was not orthogonal to the slices, a cosine

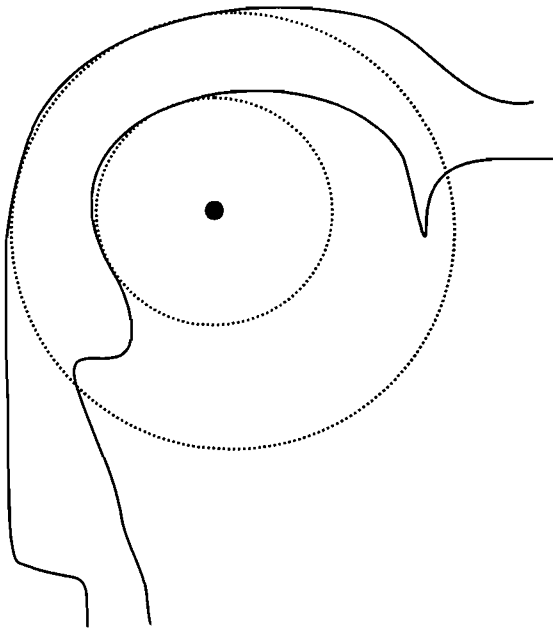

FIGURE 2.1: Vocal tract model, after Coker and Fujimura (1965). Smaller circle represents the tongue and has a moving centre with a fixed radius. The larger circle represents the hard palate and has a fixed location.

factor was multiplied with the computed area of the measured plane. This technique took advantage of the fact that bones and teeth are easily identified from midsagittal X-ray scans, whereas implementing this technique on an MR midsagittal image may raise some complications when determining bone structures that are adjacent to (air) sinuses. Since this investigation, the $\alpha\beta$-model has been further refined by several researchers for different regions of the vocal tract.

Maeda (1972) used symmetric chords to estimate sagittal distances. This method fitted circles in the tract outline, and the chord (or sagittal width line) is drawn from the point of contact between the circle and the tract outline on one side (e.g., the tongue dorsum) to the point of contact on the other side (e.g., the posterior wall of the pharynx). The midline is defined by the midpoints of the chords. It was noted, however, that discontinuities may occur at locations near side branches, as shown in Figure 2.3, which may affect the area measurements in this region.

Mermelstein (1973) provided a basis for an articulatory model of speech production. The study used midsagittal X-ray images of a subject uttering [həC−V] tokens taken at regular intervals during speech. The images were analysed individually in terms of articulatory variables to represent the static model. Temporal variations in these variables can be seen when individual images are combined to produce a relatively accurate time-varying midsagittal vocal tract outline. This information together with fundamental frequency and amplitude parameters was used to produce synthetic speech. Variables were determined by the position of the jaw, hyoid, tongue body, tongue blade,

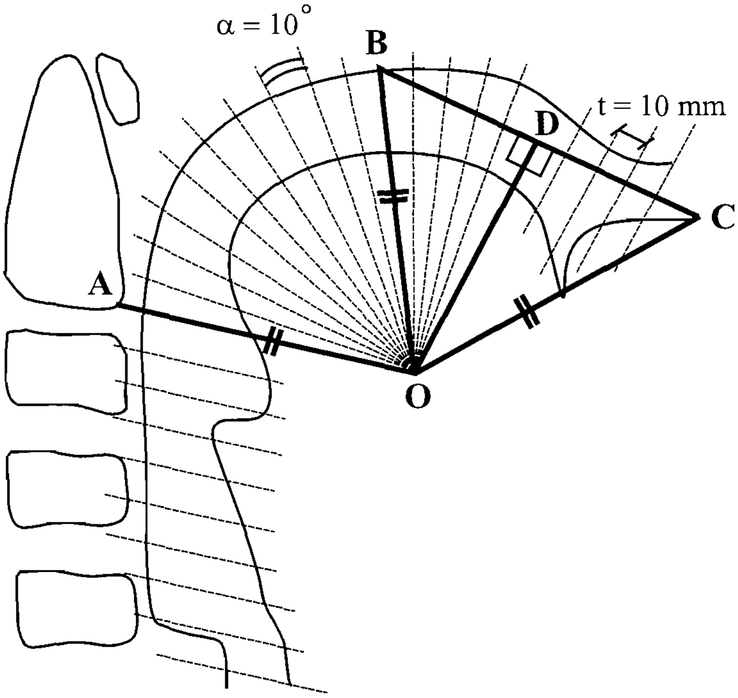

FIGURE 2.2: Vocal tract model, after Heinz and Stevens (1965). A represents the lower edge of the second cervical vertebra; B represents the posterior nasal spine; and C represents the tip of the crown of the most anterior maxillary central incisor.

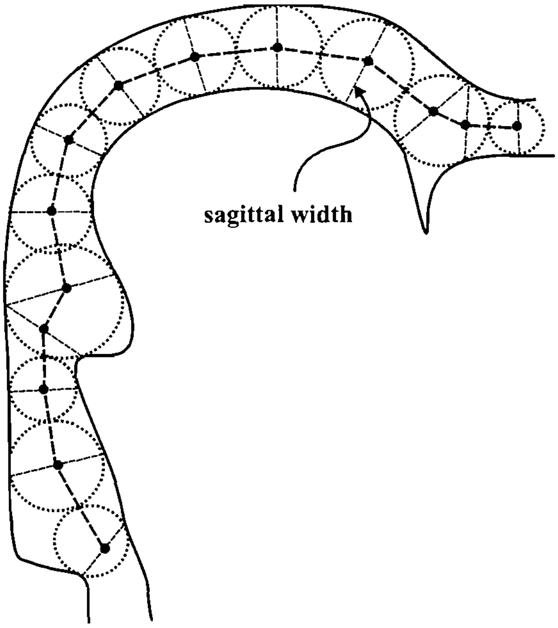

FIGURE 2.3: Using "chords" to compute sagittal widths after Maeda (1972). Goldstein (1980) noted discontinuities near side branches when using this method. The midline may be refined by placing the circles closer to each other.

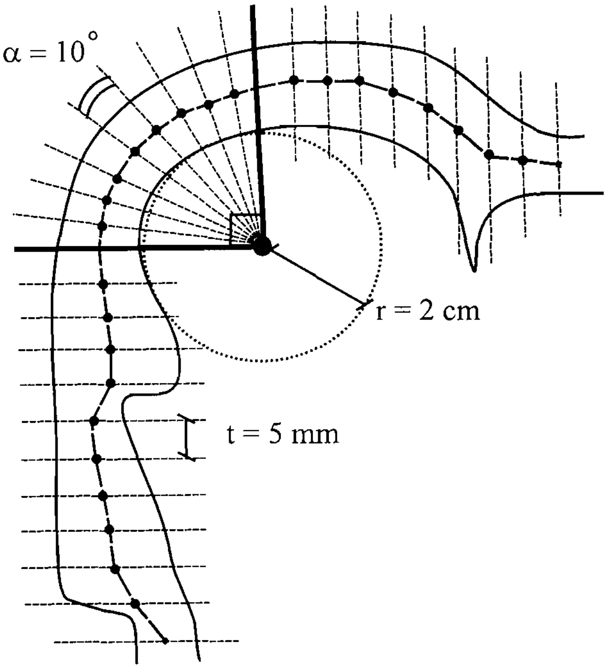

FIGURE 2.4: Vocal tract model after Mermelstein (1973). The model uses the tongue-circle concept from Coker and Fujimura (1965). The midline was derived by connecting the centre points of adjacent grid lines.

the lips and velum. The jaw and the hyoid are represented in terms of a fixed coordinate system, while the lips and tongue body are represented with respect to the moving jaw. The tongue tip position is specified relative to the tongue body. The tongue body was presented as a circle with a moving centre and a fixed radius of 2.0 cm, as suggested by Coker and Fujimura (1966). The radius of 2.0 cm was determined to be a good match across the different configurations. The tract bend was segmented radially by grid lines 10° apart converging at the centre of the tongue circle, while the remainder of the tract was segmented with parallel grid lines 5.0 mm apart. The midline of the vocal tract was determined by connecting the centre points of the sagittal slices with straight lines. The grid lines imposed on a midsagittal vocal tract outline can be seen in Figure 2.4. Following this, wave propagation at cross-sectional area, d_j, of the jth slice, was assumed to deviate from that at d_{j+1} by an angle α_j. The estimated area of d_j is calculated using $A_j = T(j, g_j) \cos \alpha_j$, where $T(j, g_j)$ is the $\alpha\beta$-transform for the pharyngeal region according to Heinz and Stevens (1964), for the oral region according to Ladefoged et al. (1971) and for the labial region according to Mermelstein et al. (1971).

An investigation by Blum (1973) on systematic methods of describing biological shapes introduced the medial axis transform shown in Figure 2.5. This technique uses circles to derive the longitudinal axis of any shape. If in our case we have a midsagittal image of the vocal tract, the circles are placed in such a way that two points on the circle are tangent to the borders of the vocal tract. The circles are placed close together so that

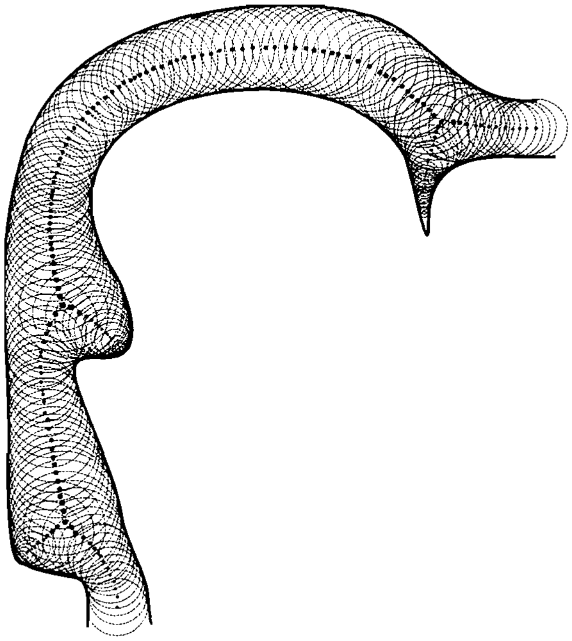

FIGURE 2.5: Blum transform, after Blum (1973). All the areas which branch off the main tract are taken into consideration. Blum (1973) found the midline of his model using chords after Maeda (1972). Miller and Fujimura (1975) and Goldstein (1980) applied this model in their investigation.

the longitudinal axis can be derived from the centre points. This technique will take into account any cavities which branch off the main tract (such as the sublingual cavity and pyriform sinuses) and therefore it can be applied to all possible configurations. In accordance with this technique, the sagittal widths are found by using symmetric chords similar to Maeda (1972). Miller and Fujimura (1975), in a study of vocal tract area functions used a simpler version of this method by placing the circles much further apart.

Goldstein (1980) developed a vocal tract growth model from infancy to adulthood to study the effects of anatomy on the production of vowels by men, women and children. A corpus of sagittal X-ray scans and data from the medical literature were used. The tongue model was similar to the one developed by Mermelstein (1973), with additional parameters for the lips, tongue tip angle, tongue tip curvature, tongue root and hyoid bone position. The midline of the vocal tract was calculated first by applying the medial axis transform from Blum (1973). Sagittal widths were measured using a radius function. This technique calculates sagittal width as twice the radius of the embedded circles. To compute the area function, the vocal tract was divided into five sections: the larynx, the pharynx, the hard palate, the alveolar ridge and the lips. Goldstein derived different equations for each section by comparing data and equations from the literature. The formant frequencies were calculated from the area functions by using a program by Henke (1966). The vocal tract parameters were altered in accordance with

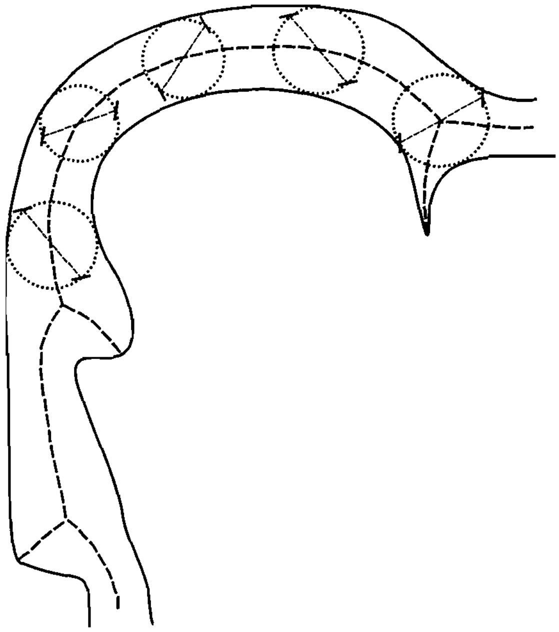

FIGURE 2.6: Vocal tract model, after Goldstein (1980). The longitudinal axis was computed using the Blum transform (Blum 1973). The sagittal widths were computed using the radius function.

the model to be simulated. Several tests were carried out using the model, including one on sudden infant death syndrome (SIDS) and the potential phonetic ability of infants. The Goldstein model is shown in Figure 2.6.

With the commercial use of MRI, researchers were able to use the technology for a further look into vocal tract configurations during speech. Because whole volume scans of the subject can be taken, and with advancements in 3-D image processing, transformation of the sagittal width data to area data was no longer necessary. However, the need for anatomical landmarks still remains, in order to allow consistent comparisons between different subjects and different tract configurations. As mentioned previously, the earliest study that used MRI for speech research is that of Baer et al. (1991), in an investigation of the vowels /a, i, u, æ/. This study used a grid system, shown in Figure 2.7, similar to that of Mermelstein (1973). However, a larger circle was employed to fit against the hard palate and posterior pharyngeal wall, which acted as the grid reference. The radius of the circle was fixed at 4.2 cm and became the centre point of the 90° radial segments, sliced at 7.5° intervals, located at the vocal tract bend. The remainder of the tract was then sliced in parallel grid lines 5 mm apart. The areas intercepted by the grid lines were calculated according to Simpson's rule and then multiplied by a scaling factor. Adjustments using linear interpolation were applied at the vocal tract bend to compensate for under-sampling (as they intersect the midline every 5.5 mm). To determine the teeth and airway boundaries, dental impression moulds of each subject made from vinyl polysiloxane were sliced in coronal sections with a

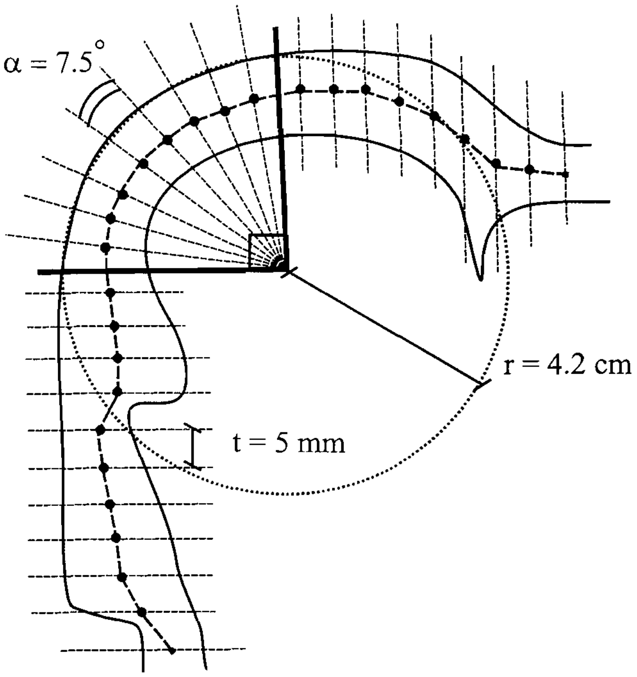

FIGURE 2.7: Vocal tract model, after Baer et al. (1991). His model was based on the grid system after Mermelstein (1973), but using the fixed larger circle concept from Coker and Fujimura (1965).

thickness of 5 mm. These slices were then digitally outlined and superimposed on the respective MR image slices identified by means of palate height, dental root structure and some other related anatomical features. A similar comprehensive study of vocal tract dimensions, this time involving fricative sounds, was conducted by Narayanan et al. (1995). Similar techniques were applied in their area function computation and the approach taken to tackle the airway and teeth boundary problems.

Beautemps et al. (1995) utilised teleradiography, a technique similar to X-ray, in their investigation of the vocal tract area function. The authors acknowledged better image acquisition methods such as MRI, but preferred acoustic recordings that were measured simultaneously with the radiographs, a feat which was not possible with the current MRI technology. The investigation focused on comparing the formants derived from area functions with those from the recorded data. The midsagittal profiles were traced by hand on transparent paper for analysis and dental impressions of the subject were also taken to increase the accuracy of the boundaries in the oral region. The grid system employed in this investigation is shown in Figure 2.8. It is made up of three different parts: parallel lines between the glottis and the low pharynx, polar lines converging to a single point between the lower pharynx and the middle of the oral cavity and parallel lines between the middle of the oral cavity to the lips. This grid system was initially proposed by Heinz and Stevens (1965). For the area function computation, a refined version of the $\alpha\beta$-model was used. Beautemps et al. (1995) altered this model by

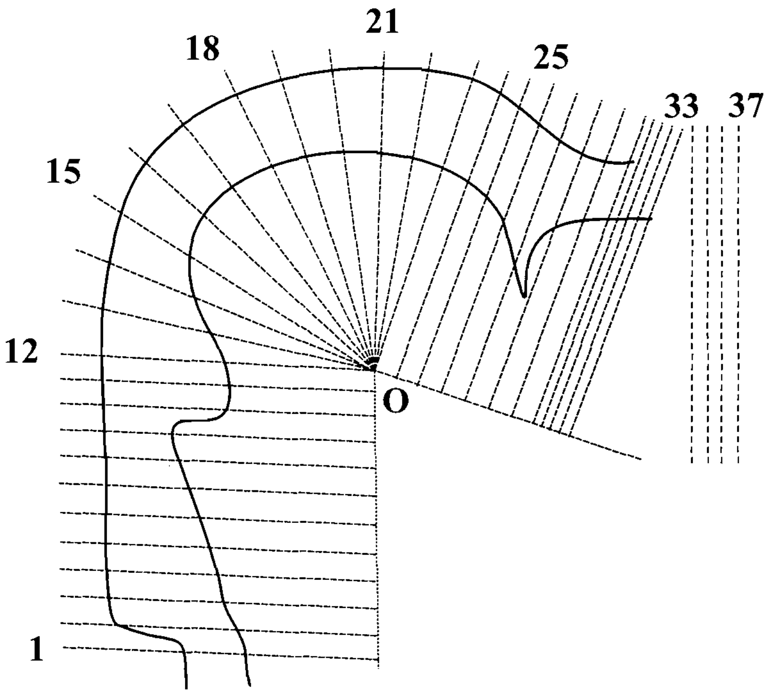

FIGURE 2.8: Grid plane as proposed by Beautemps et al. (1995). The original grid consists of planes 1–28. Note that the semi-polar planes at the vocal tract bend are the traditional widely used system from Mermelstein, Maeda, and Fujimura (1971). Engwall and Badin (1999) and Engwall and Badin (2000) employed this technique and added several more dynamic planes that can be adjusted to follow the positioning of the upper incisors (planes 29–33) and lip protrusion (planes 34–37).

keeping the β value constant, including the lip region in the calculations and excluding the laryngeal region, presenting it instead as a fixed uniform tube of 1.8 cm^2 in area and 2.0 cm in length, as suggested by Fant (1960).

Demolin et al. (1996) placed the grid planes quite differently in an investigation of the French vowels /i, e, ɛ, a, y, ϕ, œ, ɔ, o, u/ using MR images. The vocal tract was divided into three different regions, however, only parallel planes were used for each region. The horizontal and vertical planes are orthogonal to each other while the diagonal planes are angled at 45°. This model is shown in Figure 2.9. The area function was not derived; therefore, the discrepancies between the different regions were not accounted for. Comparisons were made in area and shape measurements across subjects in the same vocal tract area.

An investigation by Engwall and Badin (2000) studied coarticulatory changes of Swedish fricatives from static MR images across different vowel contexts. The subject uttered the fricatives /s, f, ɕ, ʂ, ɧ/ embedded in the vowels /a, i, o, u/. Pyriform sinuses and sublingual cavities were included in the measurements only when they were connected to the main tract on the measured plane. Dental casts and reference images were used to overcome the teeth-airway boundary problems. The slicing method used by Demolin et al. (1996)

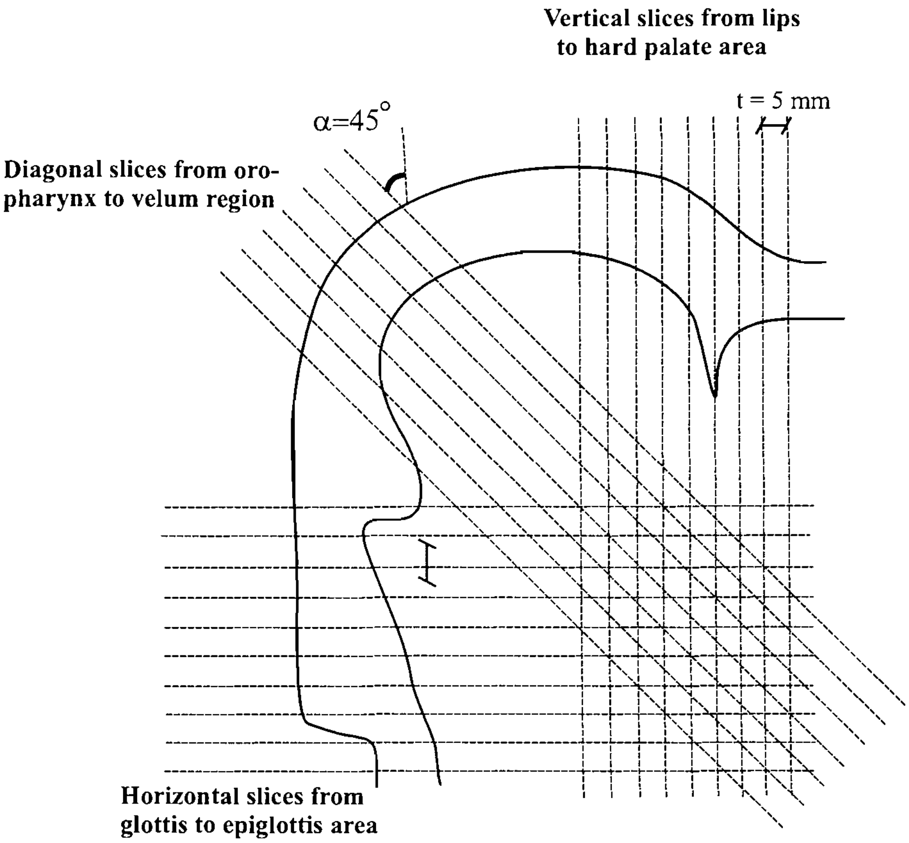
Vertical slices from lips to hard palate area
t = 5 mm
α=45°
Diagonal slices from oro-pharynx to velum region
Horizontal slices from glottis to epiglottis area

FIGURE 2.9: Vocal tract model, after Demolin et al. (1996) for a study of French vowels. The same number of slices is used for each region across subjects. Engwall and Badin (2000) adopted a similar model in an investigation of coarticulation of Swedish fricatives.

was employed here as the first part of a two-step process. The number of slices of each plane was increased to 18, such that they would overlap each other considerably. The three sets of slices were consequently processed to create three sub-tubes using the contours. The vocal tract was then reconstructed by merging the three sub-tubes together and choosing the best contours whenever an overlap occurred. The 3-D grid plane from Beautemps et al. (1995), previously shown in Figure 2.8, was used as a reference grid to reslice the reconstructed vocal tract for area function derivation.

In the above technique, the initial slicing procedure was carried out to ensure a consistent number of nodes in each sliced plane for statistical analysis. Also, additions were made to the Beautemps grid so that it was flexible and enabled the planes to move dynamically depending on the laryngeal position, tongue tip movements and lip protrusion. This was ultimately to derive parameters for the purposes of producing a 3-D "talking head".

2.2.2 Pyriform Sinuses

Because of the nature of vocal tract data in the past (which was always 1-D), the slicing methods or grid lines imposed along the longitudinal axis of the vocal tract have always been adapted to fit a midsagittal model of the vocal tract. In reality, the cross-sectional area of the vocal tract is not necessarily symmetrical, although this may not be important in parts where only a single longitudinal axis exists. However, in the laryngeal region where the tract bifurcates into three separate axes to form the pyriform sinuses, it is important to contemplate how the area in this section should be treated. Moreover, it was observed from MR images that subjects do not necessarily have symmetrical pyriform sinuses branching off the main tract (Narayanan et al. 1995). Thus, slicing the tract from a midsagittal position may result in inaccurate measurements for the posterior region of the tract.

Previous studies of how the pyriform sinuses have affected the acoustic spectra of speech have presented varying results. Fant (1960) observed that they significantly lowered the formants of vowels. Mermelstein (1967) argued they should not have any significant effect below 4.0 kHz, which was contradicted by Sundberg's (1974) observations that pyriform sinuses play a significant role in singing formants between 2.0 and 3.0 kHz. Lin (1990) observed that the acoustic effects of pyriform sinuses vary according to the vowels' articulations, the effects most prominent being on the first formant of open vowels. Fant and Båvegård (1995) observed that they significantly alter the densities of poles in the 3.0 to 5.0 kHz region and introduced a pole at about 5.2 kHz. In studies utilising MRI data, Baer et al. (1991) and Davies et al. (1993) both observed that the formants of the vowels were strongly affected when pyriform sinuses were included in the model. Baer et al. (1991) treated the pyriform sinuses as an added volume of the pharyngeal tube, while Davies et al. (1993) modelled them as side branches. In general, these studies suggest that at least for vowels, computations of the vocal tract transfer functions might be erroneous if consideration were not given to the pyriform sinuses.

One particular study by Dang and Honda (1997) compares the acoustic characteristics of the pyriform sinuses in models and humans. Data from mechanical models were used for the numerical models. In addition to these, water was injected into the pyriform sinuses of human subjects while recordings were made of sustained vowels simultaneously. Overall, the experiments produced consistent results: in cases where the pyriform sinuses were left empty, a deep trough was seen in the spectra at approximately 4.0 to 5.0 kHz which ceased to exist as the pyriform sinuses were filled. In the numerical model, comparisons were made between spectra obtained with the pyriform sinuses treated as an additional volume and treated as a separate side branch. It was observed from the spectra that the pyriform sinuses should be treated as side branches rather than additional volume because when modelled as additional volume, the spectra show incorrect behaviour in the frequency region close to the anti-resonance frequency of the branch.

It seems that there are two main reasons why the pyriform sinuses and the areas they represent have been taken so lightly in the past: first, the lack of morphological data for acoustic modelling, and secondly, their size, which is relatively small in contrast to the overall size of the vocal tract airway (Dang and Honda 1997). The importance of the pyriform sinuses in fricative modelling has yet to be determined. If the constriction is small enough that the back and front cavities are effectively acoustically decoupled during production, then it is possible that the areas of the pyriform sinuses may not be important because of almost complete cancellation of the back cavity poles. This will be discussed more in the following chapters.

2.2.3 Choice of Slicing Technique

The review of the literature has revealed that numerous techniques have been used in the past to derive the vocal tract area function. The model introduced by Coker and Fujimura (1966) provided the basis for several later models, such as the grid system employed by Mermelstein (1973) and Baer et al. (1991). In both cases, the investigators were quite vague on explaining the details of how the radius of the circle (representing the tongue ball) was determined. The authors only mentioned that the radius of 2.0 cm was found to be a good fit of the subjects' midsagittal configuration across all vowels. The grid system based on the hard palate (after Baer et al. 1991) may create non-orthogonal planes to the longitudinal axis, because it is the tongue that plays the major role in the articulation of speech. Because of this, the technique used by Mermelstein (1973) which is based on the tongue, looks more attractive. Therefore, the grid system employed by Mermelstein (1973) is investigated here.

The technique employed by Heinz and Stevens (1965) is also very popular in the literature, particularly the $\alpha\beta$-transformation. The slicing technique that was proposed, however, is highly dependent on bone structure, which may create some problems if it were to be applied on MR data. Although the second cervical vertebra and the nasal spine may be detected quite accurately with help from surrounding soft tissue, the teeth boundaries are hard to determine accurately. The $\alpha\beta$-transformation is also unnecessary, because the cross-sectional area measurements can be obtained directly from the images. The same reasoning is applied to the Beautemps grid, because the work was based on the Heinz and Stevens (1965) model. Similarly, the chords method by Maeda (1972) and the radius function does not apply with MRI, and because of this, these techniques will not be investigated here. Finally, the method from Demolin et al. (1996) is unsuitable for our purposes and was adapted by Engwall and Badin (2000) mainly for purposes of statistical modelling.

The two remaining techniques are therefore the Mermelstein and Blum area-derivation techniques. In short, they are described as follows.

1. The Mermelstein (1973) technique is based on tongue shape. The centre of origin for the semi-polar grid at the vocal tract bend is found by fitting a circle into the boundaries of the tongue surface. The areas in the anterior and posterior region of the tract are found by placing parallel grid planes at set intervals vertically and horizontally for the oral and laryngeal region respectively. The midline is found by connecting the centre points of adjacent planes to determine the cosine factor for each plane, which is then multiplied with the respective area to compensate for the plane not being orthogonal to the axis. Areas of any side branches are disregarded if they are not directly connected to the main airway on the measured grid plane.

2. The Blum (1973) transform was applied, as adapted by Goldstein (1980). The longitudinal axis is computed first by connecting the centre points of circles fitted neatly within the vocal tract boundaries. Grid planes are placed at regular intervals along the longitudinal axis at an orthogonal angle. The sublingual cavity and pyriform sinuses are presented as separate areas from the main airway because they are derived with regard to a separate axis. This will allow them to be modelled as separate branches for the acoustic predictions.

The next chapter describes the derivation of the area functions for each fricative geometry using the two area-derivation techniques.

Chapter 3

Spectral Energy in Fricatives

The first contribution of this thesis—the analysis of a fricative speech corpus—is presented in this chapter. First, a repertoire of acoustic spectra from the speech corpus needs to be derived, constantly referred to as the measured spectra. In Chapter 6, the measured spectra will be compared with the predicted spectra, which are calculated from the area functions. Because of the stochastic nature of fricative sounds, averaged power spectral density (PSD) is regarded as the most suitable approach for a study of spectral features. This is computed by averaging the spectra of eight consecutive segments of fricative steady state from a single utterance. The measured spectra were studied in terms of articulatory-acoustic relationships, spectral characteristics, such as pole and zero frequencies and relative amplitudes. Second, the acceptable frequency range of minima/maxima within the spectra is estimated by studying the variations within the eight consecutive segments of the fricative steady state. This information will help with the interpretation of the predicted spectra later on in Chapter 6. And third, a procedure is defined to quantify the influence of vocalic context on the fricative spectra. The procedure involves the use of the dynamic time warping (DTW) algorithm and mean square error (MSE) measurements on the averaged PSD spectra of the same fricative uttered in two different vowel contexts. This will be the first time that the degree of vocalic influence on each fricative will have been presented quantitatively. Subsequently, the degree to which each fricative is affected is compared to the facts presented in the literature. From here it may be determined whether the quantitative procedure is reliable and whether it can be used confidently to quantify the differences between the measured and predicted spectra (in Chapter 6).

3.1 Speech Corpus

The speech corpus consisted of spoken data, namely [vowel-fricative-vowel] utterances denoted [VFV]. The square brackets (e.g., [afa]) denote the phonetic realisation of the utterance, and forward slashes (e.g., /F/) the phonemic representation of the fricative. Generally, the phonemic representation of the fricative segment is referred to most often in the analysis. Specifically for /f/, vowels in brackets will precede the fricative to indicate the vowel context in which the fricative (audio and image) data were recorded (e.g., /(a)f/ for /f/ uttered in vowel context /a/). For the other fricatives, the preceding vowel is /a/ unless otherwise stated.

For the creation of the speech corpus, the subjects tried to replicate the sounds (in terms of stress and intonation) that were uttered during the MRI sessions as closely as possible with the assistance of a "prompt tape"; which contained the original speech utterances recorded simultaneously with the MR images. To maintain consistency with the conditions in which the image data were recorded, the speech was recorded with the subject in a supine position and the fricative segment sustained for at least two seconds. As mentioned in Chapters 1 and 2 respectively, an MRI investigation by Shadle et al. (1996) observed that the effects of vowel context on fricatives were preserved in the acoustic representation of sustained fricatives in that they exhibited features identical to the ones uttered naturally, while another investigation by Tiede et al. (2000) on contrasts in speech articulation between sitting and supine conditions observed that only articulators not actively involved during the speech utterance are affected by gravitational force when in supine position. An identical protocol used to analyse our corpus confirmed these earlier findings.

The speech utterances analysed here are [afa], [ifi], [ufu], [aθa], [asa], and [aʃa]. The fricative /f/ was also analysed in /i, u/ contexts because images were acquired for those two utterances. In the final part of this chapter, the utterances [iθi], [uθu], [isi], [usu], [iʃi] and [uʃu] along with the aforementioned utterances were analysed using DTW to measure the influence of vowel context on each fricative.

3.2 Experimental Setup

The acquisition process for the speech and image corpora is shown in Figure 3.1. Two subjects were involved, CS and PB. Only the data recorded by subject CS were analysed in this investigation. CS is a female American English speaker who is phonetically trained and has had considerable experience as a subject for fricative corpora.

The "prompt tape" was an essential tool for the creation of the speech corpus and was recorded simultaneously with the images during the image acquisition session. It contained utterances that were repeated several times before the fricative segment was finally sustained while the images were being recorded. The utterances were recorded using a Sony ES55 DAT recorder, pre-amplified and low-pass filtered at 11 kHz, before

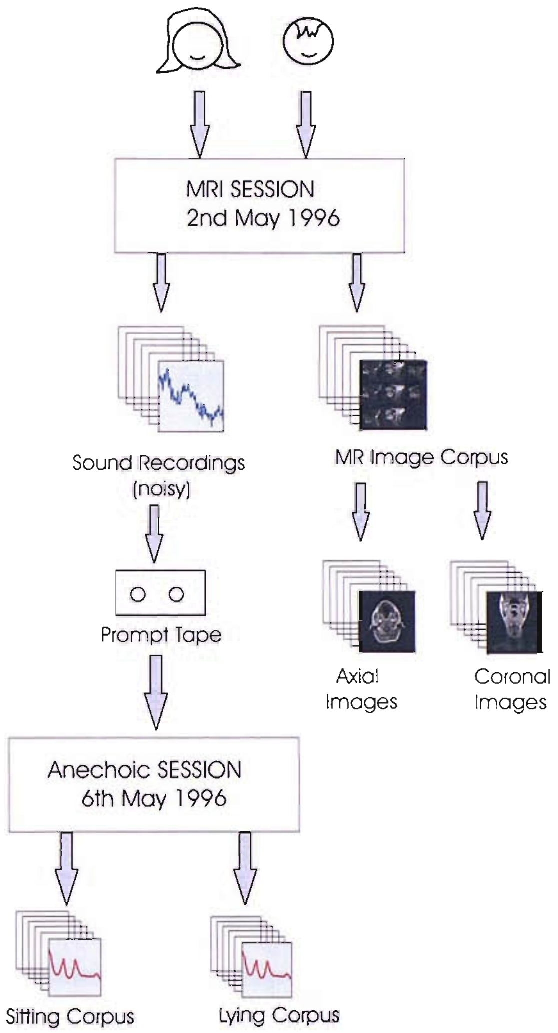

FIGURE 3.1: Corpora acquisition procedure. The resulting corpora consist of images of articulatory data and speech recordings of fricative utterances. For the speech corpus, only the sustained utterances recorded with the subject in the lying position were analysed.

being processed with an A/D converter, where they were digitised to a resolution of 16-bits with a sampling frequency of 22.05 kHz. The digitised files were edited by trimming away parts with (machine) noise (i.e., when the MRI machine was turned on).

Following the image session, on the 6th of March 1996, a recording session was held in an anechoic chamber. Before the recording, the subjects listened to the prompt tape and familiarised themselves with the stress and intonation of the utterances using Star SR Lambda Pro binaural headphones. Thereafter, the utterances were recorded first with the subjects in a sitting position and then in a lying position.

Two microphones were used for the recording. The left channel microphone was a Bruel & Kjaer 4003 model which was placed 1.0 m away from the mouth of the speaker. This microphone had a Bruel & Kjaer Type 2812 preamplifier and was connected to a Yamaha HA8 amplifier. The right channel microphone was a Sony ECM 77ES microphone, which was placed 10 cm away from the mouth of the speaker. This microphone recorded utterances onto a DAT tape using a Sony PCM-2700A recorder.

The raw DAT files were headerless, 16-bit, signed linear and had a sampling frequency of 48 kHz. A software package known as GOLDWAVE version 4.26 was used to read the binary files and convert them into `*.wav` files for processing in MATLAB.

3.3 Segmentation of Fricative Steady State

As only the fricative segment of the [VFV] utterance is used for the PSD derivation, the fricative steady state portion must be extracted first from the vocalic segment. The procedure was performed in the time-domain using the following definitions:

1. voicing offset (between the vowel and the following fricative) occurs as soon as the periodicity in the voiced section of the waveform has ceased; and
2. voicing onset (between the fricative and the following vowel) occurs as soon as the periodicity in the voiced section of the waveform appears.

For a [VF] sequence, the separation boundary was placed exactly after voice offset, although onset of noise may have gradually appeared much earlier. For a [FV] sequence, the separation boundary was placed exactly before voice onset, although noise characteristics may still be present. This procedure, although essential, does not need to be highly accurate for two reasons: locating the boundaries was done entirely for estimating the location of mid-fricative, where eight consecutive segments are placed for the PSD analysis; secondly, the segments are not on or near the transitional region of the separation boundary because only 0.4 s of fricative steady state was analysed and the sustained fricative segment had a minimum duration of 2 s.

3.4 Power Spectral Density (PSD) Analysis

The turbulence which is generated during fricative production makes the sound stochastic in nature. Because of this, an averaged PSD analysis is regarded as the best method to investigate the acoustic features. Mere application of the Fourier transform to the signal is insufficient; if each frequency component of the DFT represents the value of the estimated mean frequency, the estimate will not improve as the length of the analysed signal is increased. However, by averaging several finite-length segments of the signal, a better estimate can be found; the expected value of the average of each frequency component is then equal to the true mean value (Bendat and Piersol 1993).

By using independent (i.e., non-overlapping) segments, the variance of the estimate can be controlled by increasing the number of segments.

Eight consecutive segments, each consisting of $N = 2400$ sample points were extracted from the mid-section of the fricative steady state segmented earlier. For a 48 kHz sampling rate, each segment is 50 ms long, totalling 0.4 s in all.

The following procedure was applied on each segment, c_m, where m denotes the segment numbers 1 to 8. First, a Hamming window was applied to smooth the transitions at the start and end of the segment. Following this, the discrete Fourier transform of each segment was computed, represented by:

$$C_m[k] = \sum_{n=0}^{N-1} c_m[n] e^{-j(2\pi/N)kn} \quad k = 0, 1, ..., N-1. \tag{3.1}$$

The power spectrum, $P_m[k]$, of $C_m[k]$, is the modulus-squared of the FFT, normalised by the time resolution of the signal, N/f_s:

$$P_m[k] = \frac{f_s}{2N} |C_m[k]|^2 \quad k = 0, 1, \ldots, (\tfrac{N}{2}-1). \tag{3.2}$$

The "power per Hz" is equal to $\frac{C_m[k]}{f_s/2}$. Replacing $C_m[k]$ with this term in Equation 3.2 gives the expression for the PSD:

$$\Phi_m[k] = \frac{2}{f_s N} |C_m[k]|^2 \quad k = 0, 1, \ldots, (\tfrac{N}{2}-1). \tag{3.3}$$

Each frequency component of the eight segments was averaged to obtain the averaged PSD, given by:

$$\overline{\Phi}[k] = \frac{2}{f_s NM} \sum_{m=1}^{M} |C_m[k]|^2 \quad k = 0, 1, \ldots, (\tfrac{N}{2}-1). \tag{3.4}$$

Note that the acoustic recordings were not calibrated to provide absolute sound pressure levels.

3.5 Spectral Analysis of Voiceless Fricatives

The spectra derived using the averaged PSD computation are analysed and presented here. The features of the PSD spectra are described first and their articulatory-acoustic relationship interpreted with help from the pertinent literature. Specifically for /f/, differences in the spectra for the different vowel contexts are investigated. Following this, the variability across the eight segments within a single utterance is estimated by comparing the pole and zero frequencies of the averaged spectra with those in the individual segments.

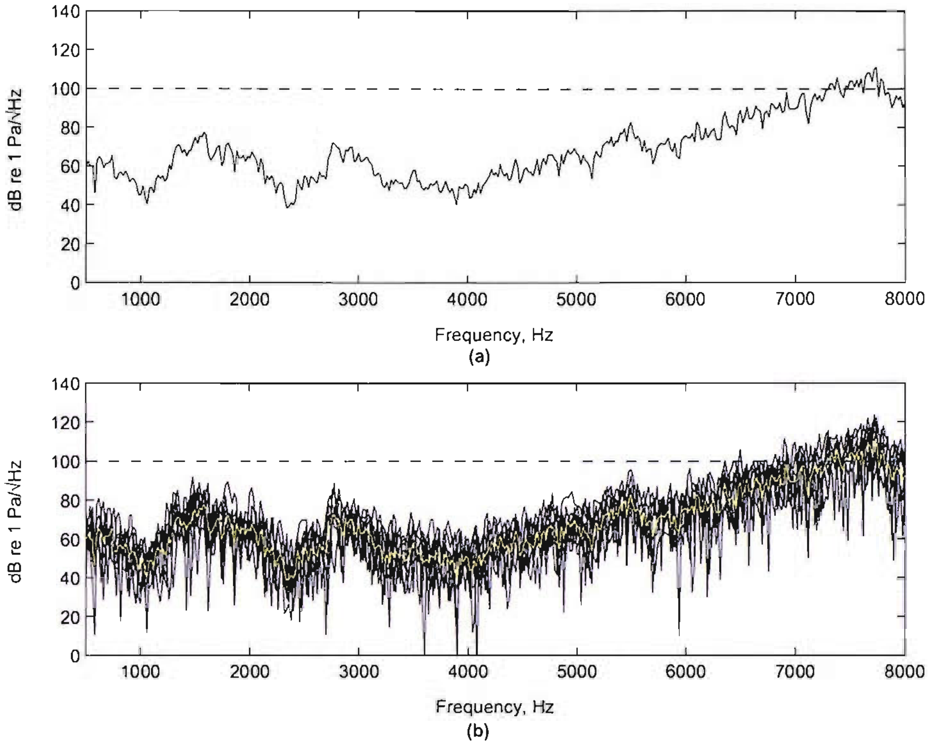

FIGURE 3.2: (a) Averaged PSD spectrum of /s/ from utterance [asa]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [asa].

3.5.1 Acoustic Features and Characteristics

Two plots of the PSD spectra are presented for each fricative. Figures 3.2, 3.3, 3.4, 3.5, 3.6 and 3.7 correspond to the fricatives /s/, /ʃ/, /(a)f/, /(i)f/, /(u)f/ and /θ/ respectively. Subplot (a) in each figure is the averaged spectral estimate of the fricative steady state and subplot (b) is an overlay of the spectra of each individual segment (a total of eight) within the fricative steady state. The PSD spectra are studied only from 0 to 8 kHz because this is the maximum frequency range of the predicted spectra. The range of the $y-$axis in the figures is fixed so that the dynamic range between the sibilant and nonsibilant fricatives is comparable. The dynamic range is an estimate of the difference between the maximum and minimum amplitudes within the spectrum. A line is drawn at 100 dB as an arbitrary threshold of significant spectral energy to help the reader interpret the spectra more easily.

The amplitude of the spectra is given in decibels, denoted as dB, which is in effect a ratio of two powers or for acoustics, two intensities, divided in the logarithmic scale and multiplied by 10 (thus the *deci* in 'decibel') (Beranek 1949). The spectra shown here were plotted as a function of frequency of 20 times the logarithm to the base 10 (the factor of 2 for the square roots of the energy), with a reference signal of 1 Pa/$\sqrt{\text{Hz}}$ as no calibrated signal was available for the recordings. Thus, the dB units do not represent

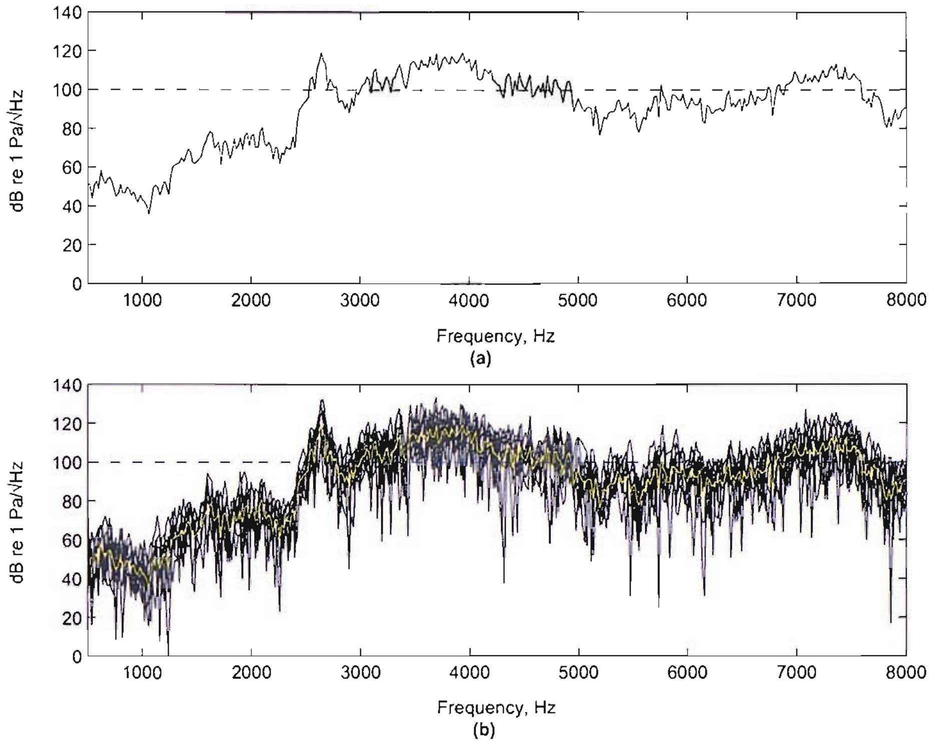

FIGURE 3.3: (a) Averaged PSD spectrum of /ʃ/ from utterance [aʃa]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [aʃa].

the ratio of the signal to a recorded referenced sound pressure. The notation on the y-axis of the spectral plots follows Kinsler et al. (1982).

3.5.2 Sibilant Fricatives

The averaged spectral estimates of the sibilant fricatives are shown in Figures 3.2 and 3.3 for /s/ and /ʃ/ respectively. The sibilant fricatives were observed to have a relatively large dynamic range compared to the nonsibilants because of the significant spectral energy in the higher frequency region. In general, the spectra of the sibilants are distinctively different from those of the front fricatives shown in Figures 3.4–3.7 and from each other.

The spectrum of /s/ can be characterised by a slow rise in magnitude above 4.0 kHz, reaching the high power threshold of 100 dB at approximately 7.0 kHz. Above 7.8 kHz the magnitude seems to decrease up to 8.0 kHz. The highest and most distinctive peak is located at approximately 7.8 kHz, while two other local maxima can be seen at 1.5 and 2.8 kHz. The peak at 1.5 kHz is relatively broad compared to the one at 2.8 kHz.

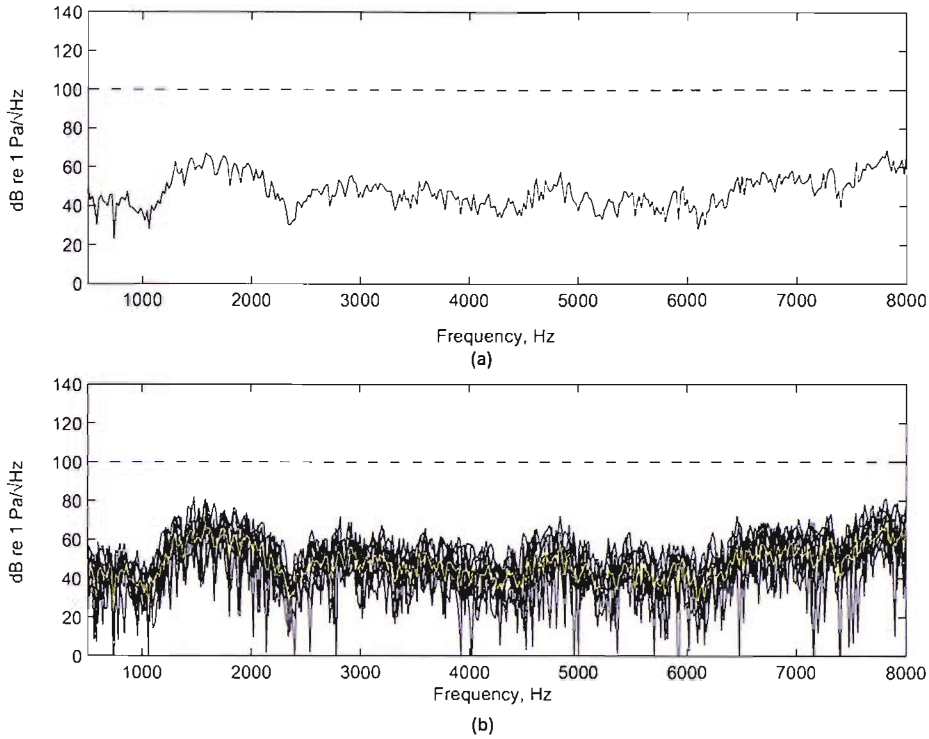

FIGURE 3.4: (a) Averaged PSD spectrum of /f/ from utterance [afa]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [afa].

The fricative /ʃ/, on the other hand, has high spectral energy from as low as 2.8 kHz (if we refer to high as being above the 100 dB threshold) up to 8.0 kHz. There is a distinctive peak—also the highest in the region—at 2.8 kHz, followed by two broad peaks, with the maxima located at approximately 3.7 and 7.4 kHz. The local minimum within this range is at approximately 5.5 kHz.

3.5.3 Nonsibilant Fricatives

The averaged PSD spectra of /f/ are shown in Figures 3.4–3.6 for vowel contexts /a, i, u/ respectively. As previously stated, this fricative is known to be the one most influenced by its vocalic environment (e.g., Stevens 1998). Stevens states this is because the tongue tip plays no specific role in forming the constriction and can be placed in several possible positions within the oral cavity during fricative production, the location being influenced by the vowel.

The gross features of all three spectra of /f/ are similar: the dynamic range is relatively small compared with the sibilants. Thus, the spectra are considered to be relatively flat. There are two local minima at approximately 1.0 and 2.4 kHz with a relatively broad peak in between. There seem to be some differences in magnitude above ~7.5 kHz: an increase in /(a)f/, relative flatness in /(i)f/ but a decrease in (u)f/.

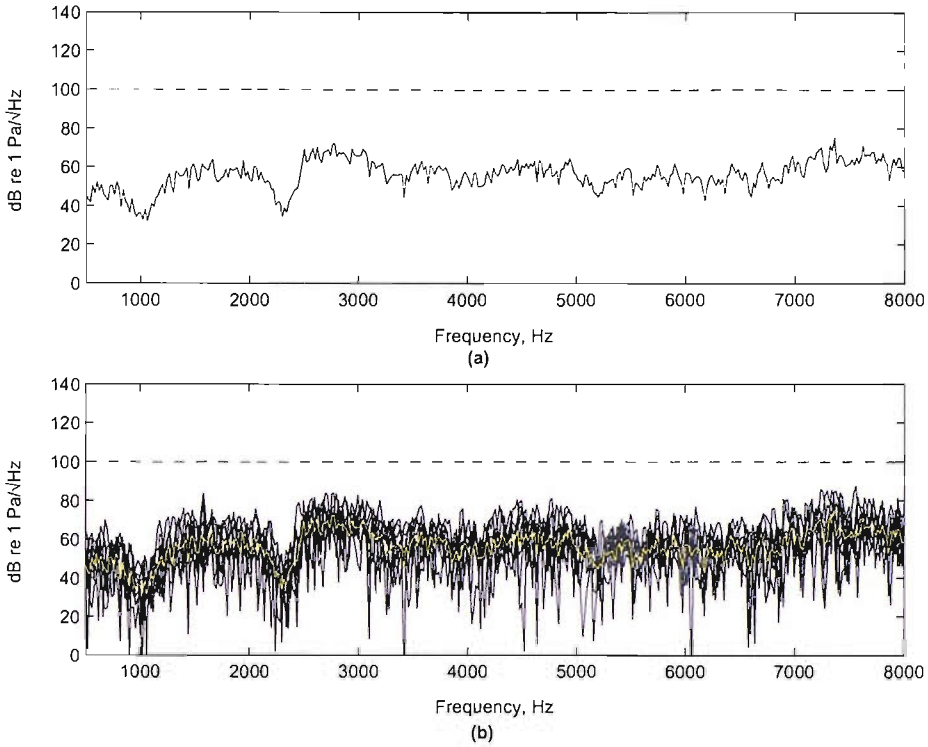

FIGURE 3.5: (a) Averaged PSD spectrum of /f/ from utterance [ifi]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [ifi].

The spectra for /θ/ are shown in Figure 3.7. Comparisons between the averaged PSD spectra of /θ/ and /f/ show similarities particularly below 7.5 kHz; the dynamic range is small. Therefore, the spectra are relatively flat and there are two local minima at 1.0 and 2.4 kHz. An investigation by Harris (1958) noted similar observations: fricative portions derived from /f/ and /θ/ cannot be used to discriminate them from each other, unlike the sibilant fricatives.

In terms of the effects of vowel context on the labiodentals, a study by Shadle et al. (1995) observed that fricatives in /a/ context produce the flattest spectra. Experiments with mechanical models indicate that this flatness may come from the effects of lip shape. The present data are in agreement with this finding; if the spectra of /f/ in the three vowel contexts are overlaid in one plot, the spectrum of /(a)f/ exhibits the smallest dynamic range.

In general, the spectral characteristics of the sustained fricatives agree well with those described in previous studies (e.g., Shadle et al. 1991; Shadle and Scully 1995; Stevens 1998; Narayanan and Alwan 2000). This is in terms of the relatively large dynamic range for the sibilants, the smaller range for the nonsibilants, the distinct features of the sibilants and the flatness and absence of well-defined peaks in the nonsibilant spectra.

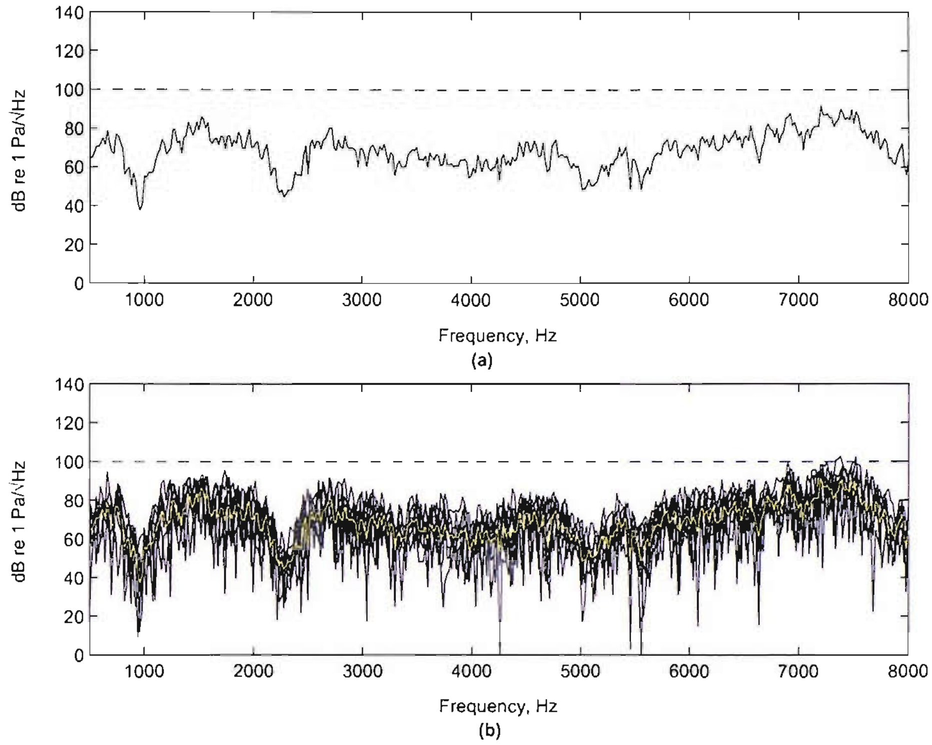

FIGURE 3.6: (a) Averaged PSD spectrum of /f/ from utterance [ufu]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [ufu].

3.5.4 Articulatory-Acoustic Relationship

This section attempts to relate the spectral characteristic with articulatory geometries with reference to previous investigations in the literature. The spectral characteristics of fricatives are somewhat complex because of the presence of source(s) above the glottis. In general, the poles in the spectra reflect the natural frequencies of the front cavity. The zeros, on the other hand, can be divided into three groups. One zero always occurs at very low frequency causing a low amplitude at low frequencies for unvoiced fricatives. One set of zeros occurs at frequencies approximately equal to the resonances of the back cavity poles, thus effectively cancelling these poles (but depending on the location of the source, the back cavity poles may not be completely cancelled). And one set of zeros occurs at frequencies related to the distance between the constriction and the source location. The most obvious spectral peaks and troughs are thus the low frequency zero, the front cavity poles and the constriction-to-source zeros (Shadle and Scully 1995).

For nonsibilants /f/ in the three vowel contexts and /θ/, characteristic peaks observed in the lower frequency range are attributed to the resonances of the back cavity and subglottal structures (Badin 1989; Stevens 1998) while the broad peak in the high frequency region, generally above 5.0 kHz, is attributed to the relatively short front cavity (Heinz and Stevens 1961). The spectra derived here agree with this: there are

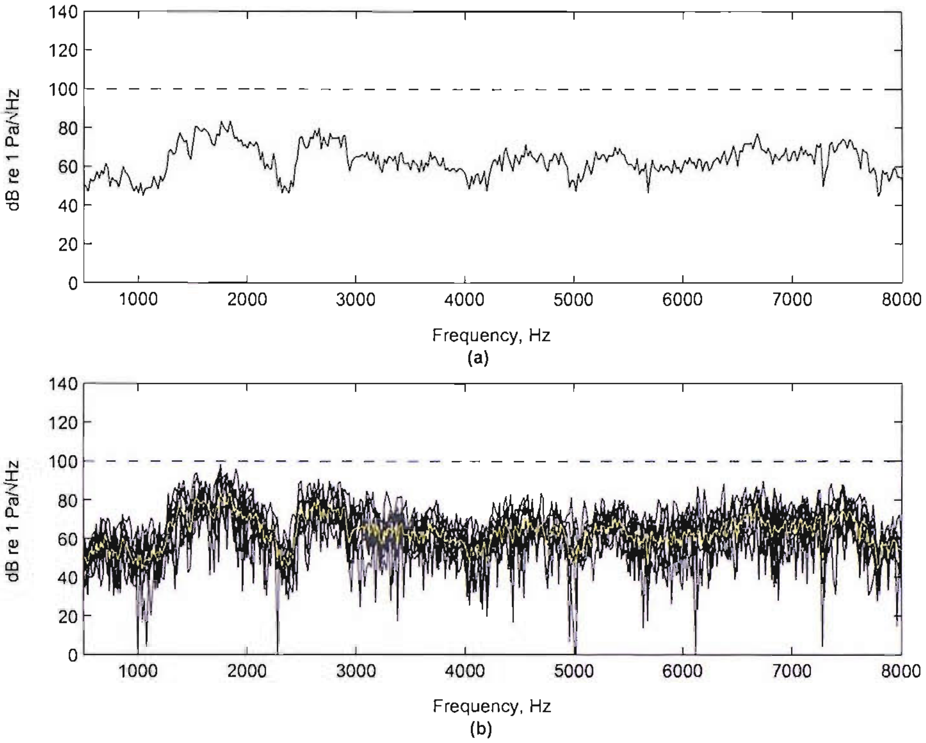

FIGURE 3.7: (a) Averaged PSD spectrum of /θ/ from utterance [aθa]. (b) PSD spectra of eight consecutive segments within fricative steady state of utterance [aθa].

local maxima (characteristic peaks) at approximately 500 Hz and 1.5 kHz. The local minima which follow each aforementioned maximum imply incomplete cancellation of the back cavity poles. The fact that these features are consistent across all the spectra of the nonsibilant fricatives indicates consistent back-cavity geometries during production. Above 5.0 kHz there are slight differences in amplitudes between the spectra, but they are relatively flat. It is possible that there are complete cancellations of pole and zero pairs occurring in this region. Above 6.0 kHz the differences between the nonsibilant spectra are more notable. This implies that geometric change to the tract mainly affects this frequency region, the geometric change naturally occurring in the front cavity region. Thus, the frequency range which has resonances affiliated to the front cavity for CS is in the vicinity of 6.0 kHz (slightly above the 5.0 kHz threshold mentioned before) which is probable since the subject is female and therefore has a vocal tract much shorter in length than the average man.

For sibilants, the distinct maximum in the region of high spectral energy is attributed to the first front cavity resonance (Heinz and Stevens 1961). Shadle et al. (1991), in a study of /ʃ/ and /s/, also obtained similar findings. Assuming this is the case, the first front cavity resonances of /s/ and /ʃ/ are located at approximately 5.5 kHz and 3.5 kHz, respectively, for this study. Below these frequencies, the power is related to the back cavity resonances, including the subglottal cavities. Narayanan and Alwan (2000) state that the strength of the spectral energy in the frequency region attributed to the back cavity—particularly close to the region of the first front cavity resonance—is related to the degree of coupling between the front and back cavities. Assuming this statement is correct, then the prominent peak at 2.8 kHz for the strident /ʃ/ denotes it as having the highest degree of coupling between the cavities across the fricatives. For the nonsibilant fricatives, it is hard to determine the level of coupling from the spectra because of the relative flatness throughput the spectra.

Notable differences in the spectra of /f/ in the three vowel contexts seem to exist above 6.0 kHz. The maximum frequency of 8.0 kHz limits the comparisons that can be made on the spectral characteristics between /f/ in the three vowel contexts, since the frequency range of interest is fixed to match that of the predicted spectra. Vowel contexts /a/ and /i/ have more similarities between them below 7.5 kHz. The spectral energy of /(u)f/, particularly in the region above 6.0 kHz, is significantly higher compared to /a, i/ context, which may have been caused by the effects of lip rounding on the radiated sound. The peaks and troughs of /(u)f/ below 3.0 kHz are lower by approximately 100 Hz from those of /a, i/, which is expected because of the increase in the effective length of the vocal tract as a result of lip protrusion. Previous studies (e.g., Shadle 1985; Narayanan and Alwan 2000) have suggested that the spectra of the nonsibilant fricatives are sensitive to slight shifts in constriction shape because of their very short front cavities and the lack of an obstacle in the path of the jet during production.

3.6 Variations Within Fricative Steady State

The lower plots for each fricative utterance in Figures 3.2–3.7 show the individual PSD spectra of the eight consecutive segments within the fricative steady state. The local minima and maxima may be distinct in the averaged estimate, but these values are clearly not consistent over all eight segments. For comparisons between the spectra, one must consider the variations brought on by the jet turbulence characteristics. Apart from this, the fluctuations in overall spectral energy between the eight segments may also be caused by fluctuations in the flowrate as CS tries to sustain the fricative. The airflow rate and consequently the resulting amplitudes of the spectra are dependent on the vocal tract geometries, such as area of constriction and length of the front cavity, as observed by Shadle (1985) using mechanical models. For an estimated 10 dB of difference in spectral energy over the eight segments, the fluctuations of the airflow rate are in the vicinity of 50 lit/s.

For the labiodentals, trough locations seem to be extensively scattered, particularly between 3.0 to 6.0 kHz, contributing to the flatness of the spectra in this region. At 1.0 and 2.4 kHz, which correspond to the frequencies of the local minima in the averaged spectra, there seem to be variations of approximately ±200 Hz in the trough locations of the individual spectra. The individual /θ/ spectra show similar attributes to those of the labiodentals.

In the case of the sibilants, the peak at 2.8 kHz for /ʃ/ seems to be relatively consistent across all segments, as is the case for the peak at 2.8 kHz for /s/, which is why the sibilant fricatives have peaks that are well-defined. The scattered troughs between 6.2 to 7.8 kHz for /ʃ/ and 6.0 to 7.8 kHz for /s/ seem to be characteristic of individual segments, as no prominent valleys are reflected in the averaged spectra. Similar to the labiodentals, the two local minima of the sibilants located at 1.0 and 2.4 kHz seem to deviate by ±200 Hz in the individual segments. This was also observed near the valley at 4.0 kHz in /s/ and the broad peak at 1.8 kHz for /ʃ/.

3.7 Effects of Vocalic Context on Acoustic Spectra

In this section, the acoustic differences in terms of magnitude and peak and trough frequencies of a fricative in two different vowel contexts are measured. To achieve this, dynamic time warping (DTW) and mean square error (MSE) measurements are used. The objective here is to determine whether the results of the measurements—which are the degree of influence of vowel context on each fricative—are in agreement with observations from previous investigations. The anticipated order of the degree of influence, starting with the most affected, is:

1. the labiodental fricative /f/ because of the insignificant role of the tongue in the formation of the constriction. The tongue tip location is highly dependent on the preceding vowel and consequently causes variation in the spectra (e.g., Shadle et al. 1996);

2. the dental fricative /θ/ and alveolar fricative /s/; /θ/ because of the very short front cavity required for its production. Thus, the spectral shape is very sensitive to small changes in the constriction shape (Shadle 1985) but less affected than /f/, because the tongue plays an important role in the formation of the constriction and /s/ because of the formation of a unique whistle-like lip geometry in the /u/ context which may possibly have introduced a second source mechanism reflected in the spectra (Shadle and Scully 1995); and,

3. the palato-alveolar /ʃ/ because it has an obstacle-type source which produces high amplitude spectra with few obvious poles and zeros, and because the longer front cavity makes the spectra less sensitive to changes in the lip geometry (Shadle 1985; Shadle 1990). The fricative /s/ also has similar characteristics (longer front cavity of the vocal tract and an obstacle type source during production) but for the /ʃ/ sound, the effects of lip rounding do not extend into the fricative.

The degree of vowel influence on /s/ and /θ/ has never been explicitly known although the most likely assumption is that the dental /θ/ is more affected than the alveolar /s/ because of the lower dynamic range of the spectrum and the sensitivity to very small changes in lip geometry. The author is aware of no cases where the effects of vowel context are measured using the quantification technique to be described here. The motivation behind this analysis is in fact related to a major part of this investigation; if the results *are* in agreement with findings from previous observations then this technique can confidently be applied to measure the differences in spectral characteristics of the measured spectra derived and previously presented in Figures 3.2–3.7, with the predicted spectra that will be derived and presented later in Chapter 6.

3.7.1 Quantification Procedure

One way to quantify the goodness of fit between two spectra is simply to take MSE measurements:

$$D(S_{\text{ref}}, S_{\text{t}}) = \frac{1}{N} \sum_{n=0}^{N-1} (S_{\text{ref}}[n] - S_{\text{t}}[n])^2, \tag{3.5}$$

where S_{ref} and S_{t} denote the reference spectrum and test spectrum respectively after expression in dB, and n is the sample number. This type of performance measure has been adopted in previous fricative investigations that compare spectra (e.g., Narayanan and Alwan 2000).

However, the computed MSE values between the two spectra can lead to wrong interpretations. For example, if the test spectrum has poles which are displaced by approximately ±20 Hz from those of the reference spectrum, the difference in the two spectra will give relatively high MSE measurements, which in turn implies that the two spectra are very different from each other. For example, consider the case where the two spectra being compared are of the same vowel /a/. However, the poles of one of the spectrum has shifted slightly as a result of, say, increase in effective length of the vocal tract. This will produce a high MSE measurement which is not desirable as it may lead to a wrong conclusion of a vowel mismatch. A solution to this problem is required.

DTW is an algorithm that warps the test spectrum temporarily to match that of the reference spectrum, so that any slight shifts in the pole and zero frequencies are accounted for. By using a "cost" matrix, the cost (or difference) between all the points of the two spectra are computed and stored. The path from point A, the top-left of the matrix, to point B, the bottom-right of the matrix, using the minimal cost possible, denoted C_{MIN}, is called the cost function, and the test spectrum is warped according to this function. The result is a reallocation of the frequencies of the main features (peaks/troughs) that were slightly mismatched in the test spectrum. If the MSE measurements are taken after DTW, the MSE values will be lower (or remain

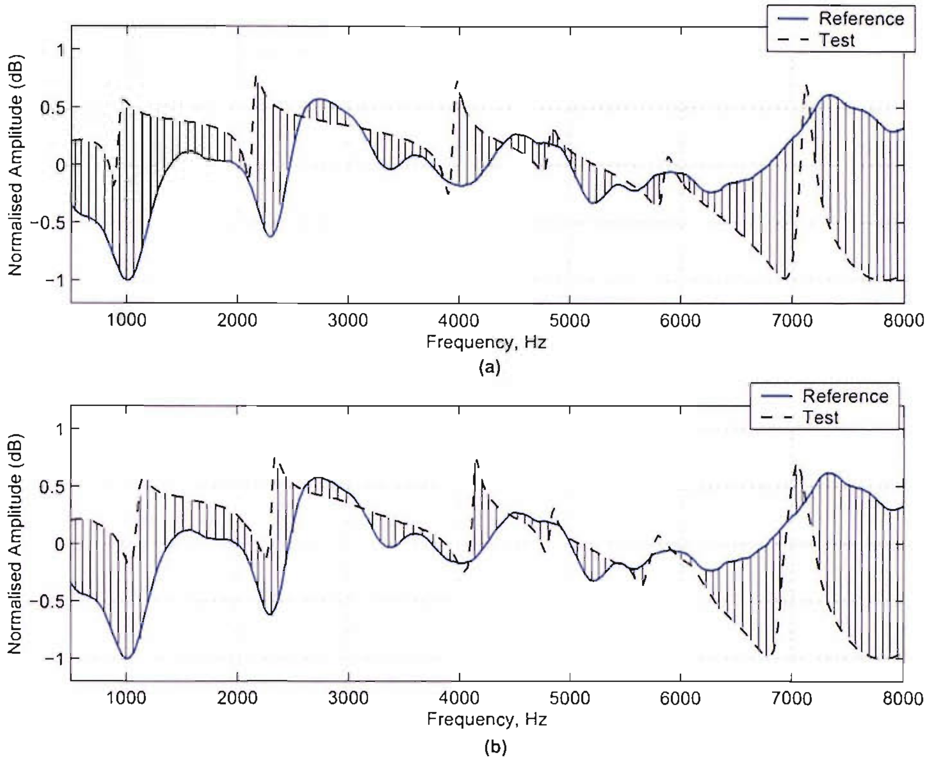

FIGURE 3.8: Application of quantification procedure on measured and predicted spectra of /(a)f/. Vertical lines represent the region of the MSE computation (a) before application of DTW and (b) after application of DTW. Note that the troughs of the test signal match better with the reference signal after warping.

the same, depending on the spectra involved). An example of the MSE measurements taken before and after application of the DTW is shown in Figure 3.8. The reference signal is the measured spectrum of /(a)f/ and the test signal is the predicted spectrum of /(a)f/. The vertical lines represent the region of MSE computation. In (a), the MSE was computed before warping the test spectrum and in (b), the MSE was computed after warping the test spectrum. Note how the troughs of the test (predicted) spectrum at 2.3 kHz and 3.9 kHz have shifted to match those in the reference (measured) spectrum. Consequently, the MSE measurement of (b) is less than the MSE value of (a).

Application of the DTW first and taking MSE measurements second will acknowledge any slight mismatches in the peaks and troughs of the two spectra. It emerges that the computed C_{MIN} is in fact proportional to the MSE values measured after warping. If this is the case, then we need only the C_{MIN} *or* MSE values to quantify the difference between the two spectra. But consider the case where the two spectra that are compared are significantly different from each other. An example of such a scenario can be seen in Figure 6.4 in Chapter 6. Because of the complete mismatch between the two spectra, very little or no warping at all was applied on the test spectrum; thus the computed minimal cost is very low. Low minimal costs (or low MSE values, since the two are proportional) imply that the two spectra are a good match, but from mere observation alone one knows that this is certainly not the case. To obtain a good measure of the degree of match between two spectra, the MSE measurements made before and after DTW must be considered. Thus we define the degree of match between two spectra, D_m, to be given by:

$$D_m = \frac{E_B - E_A}{E_B} \times 100 \tag{3.6}$$

where E_B and E_A refers to the computed MSE value before and after application of DTW respectively. A high D_m value indicates that the test spectrum is a good match to the reference spectrum in terms of peak/trough frequencies. Conversely, a low D_m value indicates that the test spectrum is a poor match to the reference spectrum or completely different to the test spectrum.

Several examples are presented here to support the definition of the degree of match, D_m. Consider the case where the two spectra that are being compared are nearly identical to each other except for a small shift of say, 1.0 Hz in the test spectrum. An example is shown in Figure 3.9 with the corresponding MSE and D_m values shown in the second row of Table 3.1. The first MSE measurement before DTW is relatively high (considering the similarity between the spectra) because of the shift but the second MSE after DTW is considerably lower. The final result is a D_m value of 99%, which allows the correct conclusion that the two spectra were a good match. Now consider another case where the two signals do not show any similarities at all in the peak and trough frequencies but they do have similar amplitude levels. An example of this scenario is shown in Figure 3.10 with corresponding values presented in the third row of Table 3.1. The MSE computed before warp is relatively low (considering the mismatch between the spectra). After DTW, the MSE was only slightly reduced because the DTW algorithm was unable to find a better minimal path. This resulted in a D_m value of 5%, which allows the correct conclusion that the spectra were a poor match. Finally, an example is given where the two spectra are completely different, both in terms of peak and trough frequencies and in terms of amplitude levels. An example is given in Figure 3.11 with corresponding MSE and D_m values given in the fourth row of Table 3.1. As can be seen here, the MSE measurements before and after warping were high which resulted in a degree of match of only 5%, again, indicating correctly that the two spectra were completely mismatched.

For the application stated here, the D_m value allows quantification of the degree of the vocalic influence on the fricative. The relationship is in fact inversely proportional; a high D_m indicates a good match between the compared spectra, therefore the less affected the fricative is by vocalic context. Conversely, a low D_m indicates a poor match between the compared spectra; thus, the more affected the fricative is by vocalic context.

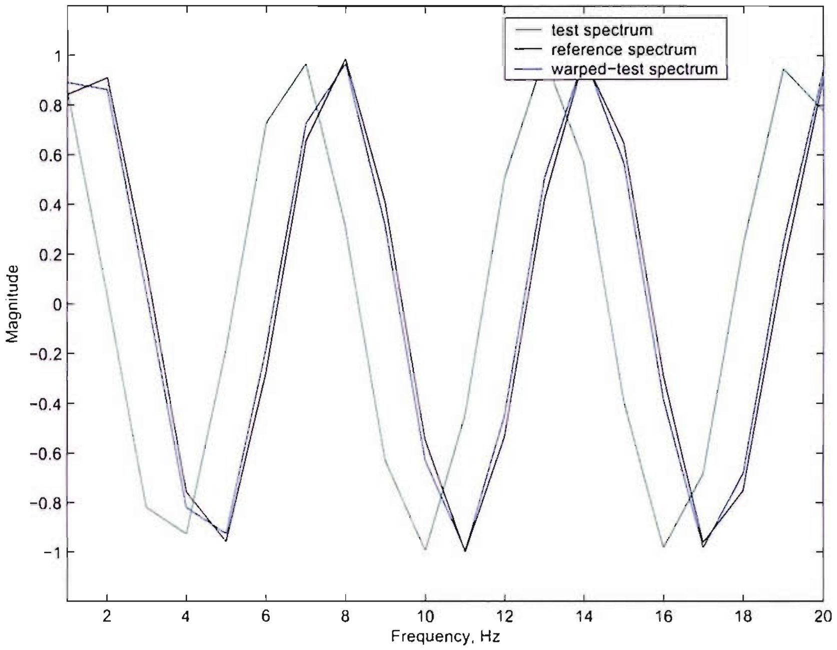

FIGURE 3.9: Example 1 for D_m computation; two nearly identical spectra, one spectrum shifted.

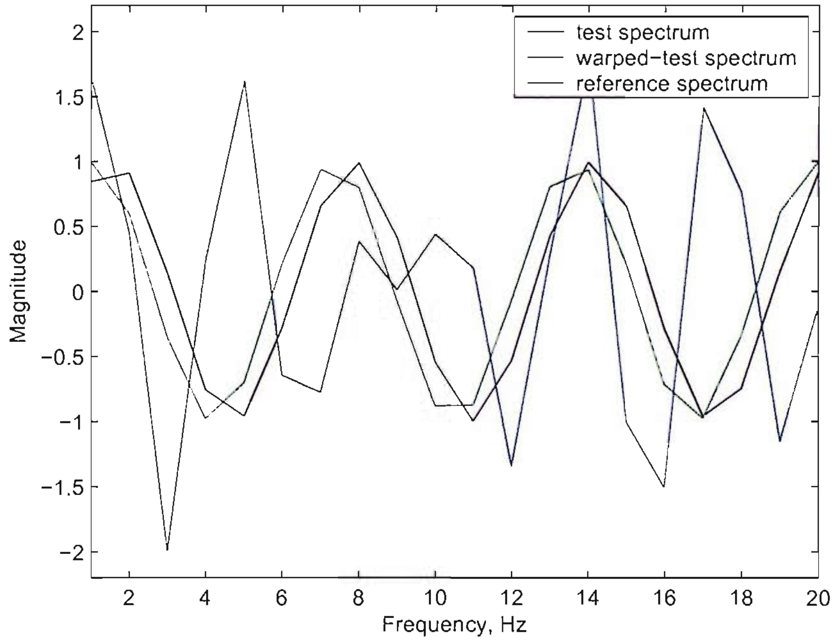

FIGURE 3.10: Example 2 for D_m computation; similar amplitude-level spectra.

Example no.	Description of signals	E_B %	E_A %	D_m %
1	Nearly identical spectra, one shifted	44 (low)	1 (low)	99 (high)
2	Similar amplitude-level spectra	43 (low)	41 (low)	5 (low)
3	Two completely different spectra	88 (high)	84 (high)	5 (low)

TABLE 3.1: Examples for degree of match, D_m. The terms E_A and E_B denote the MSE measurements before and after DTW respectively and D_m denotes the computed degree of match values. The measurements given here are relative to each other. Note that low MSE values do not necessarily indicate a good match between the spectra, as can be seen from Figures 3.9 to 3.11.

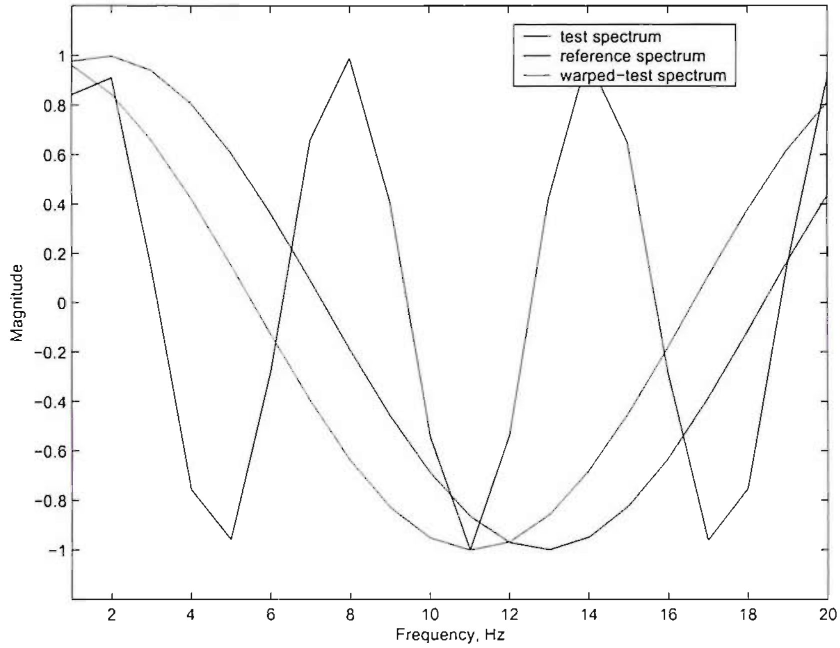

FIGURE 3.11: Example 3 for D_m computation: two completely different spectra

Therefore, we define the degree of vocalic influence on the fricative, D_v, as:

$$D_v = 100 - D_m, \tag{3.7}$$

where D_m is the degree of match between two spectra as defined in Equation 3.6.

In the following section, the quantification procedure was applied between two spectra of the same fricative in different vowel contexts, the vowels being /a, i, u/. To summarise, for two spectra of the same fricative in two different vowel contexts, the following procedure was applied:

1. the MSE between the reference and the test spectra is measured to obtain E_B;
2. the DTW is applied and consequently the test spectrum is warped;

Fricative	/f/				/θ/			
	E_B	E_A	D_m (%)	D_v (%)	E_B	E_A	D_m (%)	D_v (%)
$\overline{\Phi}_{[a]}$ and $\overline{\Phi}_{[i]}$	145.10	108.63	25.13	74.87	57.16	22.20	61.16	38.84
$\overline{\Phi}_{[i]}$ and $\overline{\Phi}_{[u]}$	516.90	422.31	18.30	81.70	269.49	201.45	25.25	74.75
$\overline{\Phi}_{[u]}$ and $\overline{\Phi}_{[a]}$	226.56	165.97	26.74	73.26	342.79	254.36	25.80	74.20
Average			21.57	78.43			28.60	71.40
Fricative	/ s/				/ʃ/			
	E_B	E_A	D_m (%)	D_v (%)	E_B	E_A	D_m (%)	D_v (%)
$\overline{\Phi}_{[a]}$ and $\overline{\Phi}_{[i]}$	62.13	30.97	50.15	49.85	178.74	110.44	38.21	61.79
$\overline{\Phi}_{[i]}$ and $\overline{\Phi}_{[u]}$	305.60	234.17	23.37	76.63	276.43	183.28	33.70	66.30
$\overline{\Phi}_{[u]}$ and $\overline{\Phi}_{[a]}$	288.91	230.97	20.05	79.95	106.40	65.80	38.16	64.02
Average			24.45	75.55			35.98	64.02

TABLE 3.2: Results of quantification procedure for fricative spectra in different vowel contexts. E_B and E_A denote the MSE measurements taken before and after DTW respectively. D_m and D_v denote the degree of match between the spectra and the degree of vocalic influence, respectively.

3. the MSE between the reference and the warped test spectra is measured to obtain E_A; and

4. the degree of match, D_m, is computed and subsequently the degree of vocalic influence on the fricative, D_v.

3.7.2 Results and Discussion: Degree of Vocalic Influence

Table 3.2 shows the results of the quantification procedure for the nonsibilant and sibilant fricatives respectively. The first column lists the vowel contexts of the two spectra that were compared. For example, $\overline{\Phi}_{[a]}$ and $\overline{\Phi}_{[i]}$ for /ʃ/ means that the averaged PSD spectra of /ʃ/ in vowel context /a/ is compared to the averaged PSD spectra of /ʃ/ in vowel context /i/. The degree of match between the two spectra, D_m is given in columns 4 and 8 and the degree of vocalic influence, D_v is given in columns 5 and 9. The averaged D_m and D_v values are given in rows 6 and 12.

It was observed that the rank order of the degree of vowel context effect on the fricative, starting with the most affected is:

1. the nonsibilant /f/;

2. the sibilant /s/;

3. the nonsibilant /θ/; and

4. the sibilant /ʃ/.

The order is in agreement with the anticipated order previously given in the first part of this section. The nonsibilant /f/ is most affected by vowel context with an average D_v value of 78%. The degree of match is relatively similar for comparisons involving /a/ contexts (25% for /a/–/i/ and 26% for /u/–/a/), while the comparison between vowel contexts /i/–/u/ show the lowest degree of match for this fricative (18%). Apart from tongue location, this may potentially be caused by different lips geometries, rounding for the /u/ context and retraction for /i/. Because of the shorter front cavity and the absence of an obstacle in the path of the jet turbulence, this fricative is highly sensitive to changes in constriction shape and the location of the frication surface (Shadle 1985).

The sibilant /s/ is second most affected by vocalic context with an average D_v value of 76%. The comparisons involving /u/ context show the poorest degree of match (23% for /i/–/u/ and 20% for /u/–/a/), lowering the overall D_m average, while comparison in /i/–/a/ contexts shows a higher degree of match (50%). The poor mismatches in /u/ context support the findings of Shadle and Scully (1995) who suggested that the effects of lip rounding extends into the fricative creating a second source mechanism brought on by the whistle-like configuration of the lips.

The nonsibilant /θ/ is the third most affected by vocalic context with a D_v value of 71%. In general, the measured D_m values are similar to that of /s/, smaller values for comparisons involving /u/ context (25% for both /i/–/u/ and /u/–/a/) and a higher value for /a/–/i/ contexts (61%). The variations in this fricative, however, are more likely to do with the sensitivity to lip shape, similar to /f/. Finally, the sibilant /ʃ/ is least affected by vowel context with a significantly lower D_v value of 64%. Comparisons involving /a/ context show similar values (38% for /a/–/i/ and /u/–/a/) while the comparison in /i/–/u/ context show a lower value (33%). For this fricative, the tongue plays a major role in forming the narrow channel for the constriction and the teeth act as an obstacle for the sound generation mechanism. The teeth are fixed, which may explain why this fricative is least affected by vowel context.

The results here are interesting, because although it is known that some fricative spectra show more variation than others as a result of vowel contexts, the degree of vocalic influence on each fricative has never been quantified. Improved understanding might be obtained by comparisons using a greater number of samples for each fricative. For now, the fact that the results are in agreement with observations from previous studies indicates that we can apply the quantification procedure to measure the differences between the measured and predicted spectra in Chapter 6 with some confidence.

3.8 Summary and Conclusions

This chapter describes the investigation of the spectral energy of sustained fricatives in different vowel contexts. The main objective was to derive a set of fricative spectra that can be used for comparison with the predicted spectra in Chapter 6. Because of the stochastic nature of fricative sounds, the averaged PSD is the best way to study the acoustic features. It was computed over eight consecutive segments within the fricative steady state segment and averaged over each frequency component. The spectral characteristics of the fricatives were compared and discussed in terms of the acoustic-articulatory relationships with reference to observations from previous investigations. The variations within each (consecutive) segment in fricative steady-state were also investigated to determine the variance of the frequency of the maxima/minima in the averaged PSD. In the final section, the effects of vowel context on the fricative spectra were quantified using DTW and MSE measurements.

The study of the spectral features shows that the nonsibilant fricatives /f/ and /θ/ have relatively low dynamic range in comparison to the sibilants /s/ and /ʃ/. Because of this, the spectra of the nonsibilants are quite flat, characterised by broad peaks. In contrast, the sibilant spectra have a large dynamic range because of the high spectral energy particularly in the higher frequencies. For the nonsibilant fricatives, /θ/ shows similar features to the /f/ spectra, making it hard to differentiate between them. This supports the finding of Harris (1958). On the other hand, the sibilants have unique features in terms of the distribution of the spectral energy and local maxima/minima. In general, however, two local minima can be seen in all the spectra at approximately 1.0 and 2.4 kHz, which might be attributed to back cavity resonances.

The effects of vowel context on the spectra of /f/ are not clear because of the relative flatness and the absence of well-defined peaks. Significant differences between the three spectra were observed to occur above 7.5 kHz, but the maximum frequency of the spectra was fixed at 8.0 kHz, which limits the comparisons. It was observed that the spectra of [ufu] had higher energy levels than [afa] and [ifi], possibly an effect of lip rounding as well as tongue location.

The PSD spectra of the eight consecutive segments within a single fricative utterance were overlaid on a single graph so that the variations in the frequencies of the local maxima/minima could be visualised. It is speculated that the differences in the overall amplitude levels of the individual spectra may be attributed to the variations in the noise characteristics and possibly fluctuations in airflow rates made involuntarily by the subject as she attempted to sustain the fricative. It was observed that the distinct minima/maxima locations of the averaged spectra were dispersed by approximately ±200 Hz in the individual spectra. This finding enables a conclusion that the predicted spectra is still a good match to the measured counterpart if the local maxima/minima frequencies are within ±200 Hz of the expected frequency.

Finally, the quantification procedure shows that the degree of vocalic influence on each fricative, from most affected to least affected, are: /f/, /θ/, /s/ and /ʃ/. This order agrees well with findings from the literature, although the degree of vocalic influence in /s/ and /θ/ has never been explicitly established. The labiodental /f/ was observed to be the most highly influenced by vowel context, perhaps because of the differences in tongue location, while the fricative /s/ shows variations (between spectra of different vowel contexts) which may be because lip rounding extends into the fricative. Thus, comparisons with spectra in the /u/ context show the poorest matches. The fricative /θ/ is sensitive to constriction shape because of the shorter front cavity, but is less affected by vowel context than /f/ because of the major role played by the tongue in forming the constriction. Finally, the fricative /ʃ/ shows the best matches between the spectra, indicating that it is least affected by vowel context. This is probably because the tongue plays a major role in forming the constriction. The front cavity is longer making it less sensitive to lip shape and the teeth act as an obstacle for the source mechanism resulting in high-amplitude spectra with few obvious peaks/troughs. The good agreement between the outcome of this procedure and the literature—indicated by these results—allows us to apply this measurement technique to quantify the differences between the measured and predicted spectra in Chapter 6 with some confidence. Before that, Chapters 4 and 5 deal with derivation of the area functions and introduce the reader to the acoustic prediction software, ACTRA, respectively.

Chapter 4

MRI Data Analysis: Area Function

The previous chapter presented an analysis of the spectral properties of the speech corpus and defined a procedure to quantify the degree of match between two spectra. In this chapter, the second contribution of this thesis, the analysis of the image corpus, is presented. The objective was to derive a set of area functions from the volumetric magnetic resonance (MR) images of the vocal tract. In Chapter 6, these area functions will be used to calculate the predicted spectra. The area functions are extracted from the images using the Blum transform and the Mermelstein technique previously described in Chapter 2. The major part of this chapter will cover the implementation of the area-derivation methodology on the three-dimensional (3-D) images and the solution that was adopted to overcome the teeth- and airway-boundary problem. Since two area-derivation techniques will be used, two sets of area functions will be obtained for each fricative. The discussion will be based mainly on differences between area functions for the same fricative, since at this stage we cannot speculate which area-derivation technique will produce the best predicted spectra and consequently synthetic sounds. For the labiodental /f/ (previously observed to be the fricative most affected by vowel context), the area functions of the fricative uttered in different vowel contexts will also be compared.

4.1 Image Corpus

Acquisition of the image corpus has been described briefly in Section 3.2 (Figure 3.1, page 37). Magnetic resonance imaging (MRI), the technology already described in Chapter 2, captured the images through the spin-echo technique for high resolution. The echo time, T_E, and repetition time, T_R, two important parameters for MR imaging, were 15 ms and 920 ms respectively. The images were recorded in two sessions: the

Fricative	Session date	Orientation	Number of slices	FOV (mm)	Slice thickness (mm)
/(a)f/	24/2/96	Coronal	30	258.1	4
/(a)f/	24/2/96	Axial	30	258.1	5
/(i)f/	2/3/96	Coronal	30	266.7	4
/(i)f/	2/3/96	Axial	25	266.7	5
/(u)f/	2/3/96	Coronal	30	266.7	4
/(u)f/	2/3/96	Axial	25	266.7	5
/(a)θ/	24/2/96	Coronal	30	258.1	4
/(a)θ/	24/2/96	Axial	30	258.1	5
/(a)s/	2/3/96	Coronal	30	266.7	4
/(a)s/	2/3/96	Axial	25	266.7	5
/(a)ʃ/	24/2/96	Coronal	30	258.1	4
/(a)ʃ/	24/2/96	Axial	30	258.1	5

TABLE 4.1: Details of the image corpus that was used in this investigation. Two separate sessions were held to acquire the images. The fricative segment was sustained for a minimum of four minutes while the images were being taken.

first session was held on the 24th of February 1996 for the fricatives /(a)f, ʃ, θ/ and the second session was held on the 2nd of March 1996 for the fricatives /s, (i)f, (u)f/. As previously mentioned in Chapter 3, the vowel in brackets preceding the labiodental /f/ indicates the vowel context of the fricative, while for the other fricatives, the vowel context is /a/ unless stated otherwise. The field of view (FOV) parameters for the two sessions were 258.1 mm and 266.7 mm respectively.

Speaker CS sustained the fricative segment for a minimum of four minutes for a complete volume scan taken in axial and coronal directions. The thickness of the slices was 5.0 mm in the axial direction and 4.0 mm in the coronal direction. A full volume scan consists of 25–30 images, ranging from the larynx area up to the hard palate for the axial scans and from the lips to the back of the pharyngeal wall for the coronal scans. The details given here are listed in Table 4.1.

4.1.1 Preliminary Processing of Images

Preliminary processing of the MR images utilises a medical image processing software known as 3D-Doctor. Raw MR image data were first converted into TIF format using the software. An example of a set of axial MR images of the subject sustaining /(a)f/ is shown in montage view in Figure 4.1.

In the image slices in Figure 4.1, the whitest areas indicate regions with the highest concentration of hydrogen, such as fatty tissue and bone marrow. The muscles and connective tissues appear in varying shades of grey, depending on hydrogen levels, while the darkest regions show the air spaces, such as the vocal tract airway and sinuses and

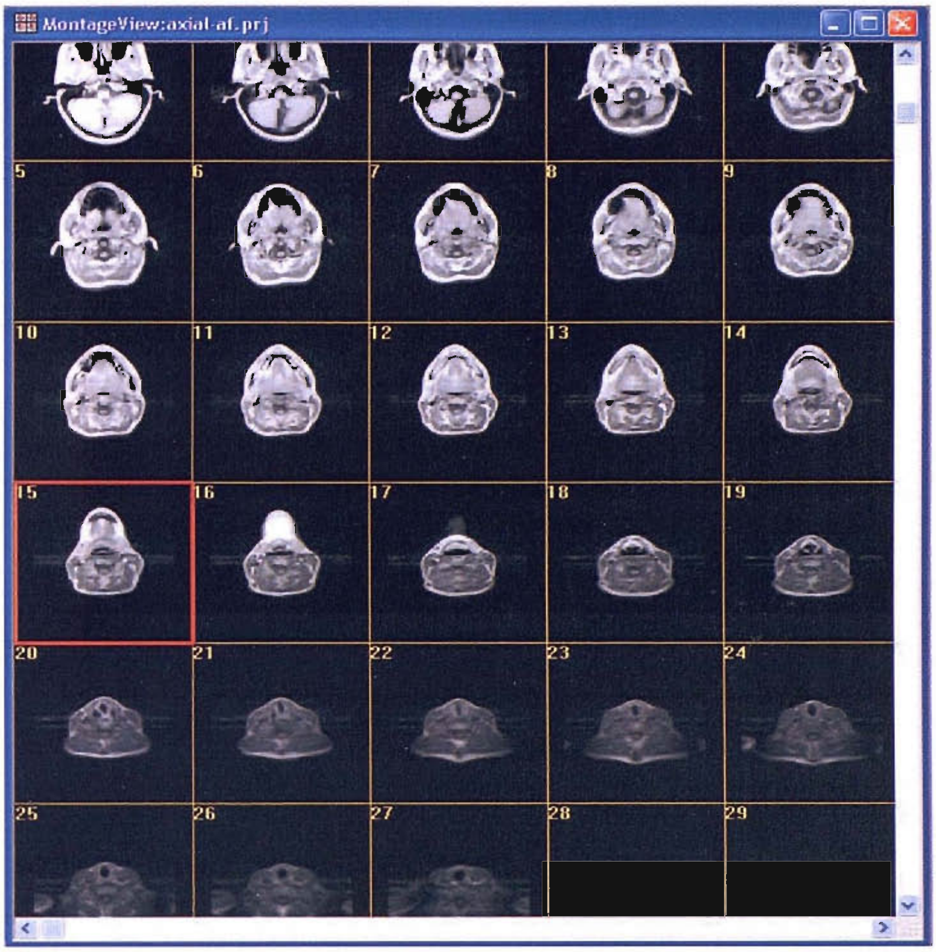

FIGURE 4.1: Axial MR images in montage view. The raw images were first converted into `*.tif` format for processing. The example shown here are axial images of subject CS sustaining the fricative /(a)ʃ/.

calcified structures such as bone and teeth, where the concentration of hydrogen is very low.

The next step involves tracing the outline (or border) of the subject's profile and vocal tract airway. The boundaries of the vocal tract airway are important for the derivation of the area function, while the boundaries of subject CS's profile are important for aligning the coronal and axial images. Apart from the teeth-airway boundaries, which will be explained in more detail in the next section, the boundary outlining process was done automatically using the segmenting function in 3D-Doctor. For accuracy, the boundaries were further inspected by hand to remove or add any boundary nodes ignored or overlooked by the function and to delete any unnecessary boundaries picked up by the segmenting function. At locations where the boundaries between the airway and surrounding tissues were vague, the threshold of the contrast function was increased to assist in boundary determination. This adjustment was necessary as the image slices were relatively thick (i.e., 4.0 and 5.0 mm) and changes in the vocal tract shape do

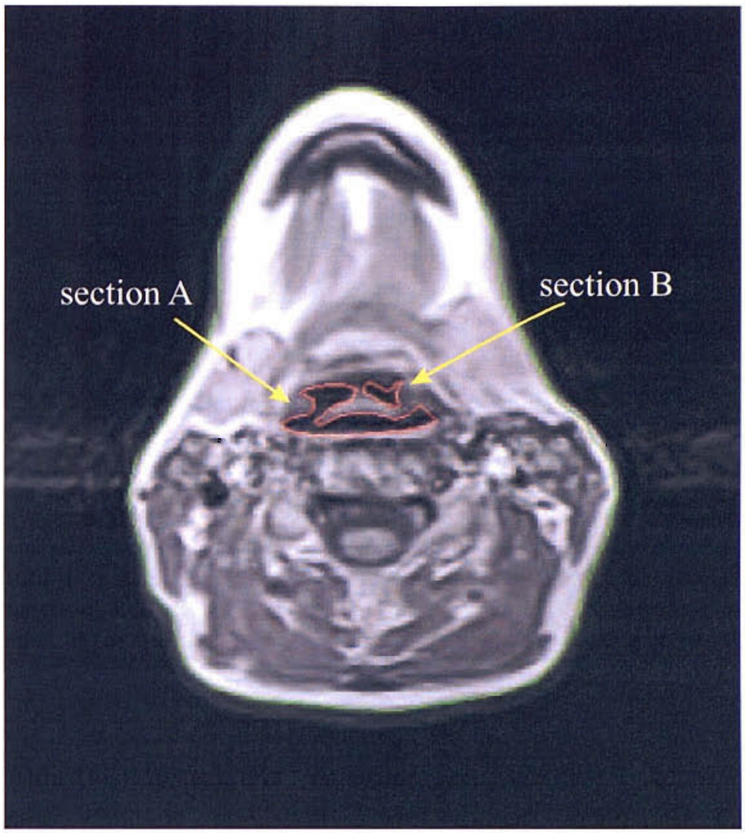

FIGURE 4.2: Axial image slice with outlined boundaries. This example also shows the treatment of side cavities for the Mermelstein and Blum techniques. Only section *A* is included in the Mermelstein area functions while both section *A* and *B* are included in the Blum area functions.

occur within this interval. For consistency, only completely black areas were included as an airway, while dark-greyish areas were ignored. An example of an image slice with completed boundaries is shown in Figure 4.2.

After the boundaries of the vocal tract airway and the subject's profile had been determined, the coronal slices and the axial slices were rendered (separately) to form two complete 3-D models of the head and tract. An example of a 3-D model made from axial slices in two different angles is shown in Figure 4.3.

The following section describes how the boundaries between the teeth and vocal tract airway were determined in the coronal slices and then how the two area-derivation techniques were implemented on the 3-D models. Going back to the boundary-derivation procedure, extra attention was given to the boundaries in the region from the vocal tract bend towards the laryngeal region for the axial scans, while for the coronal images, emphasis was given from the lip region to the vocal tract bend. This is because area measurements of the anterior and posterior region of the vocal tract were taken

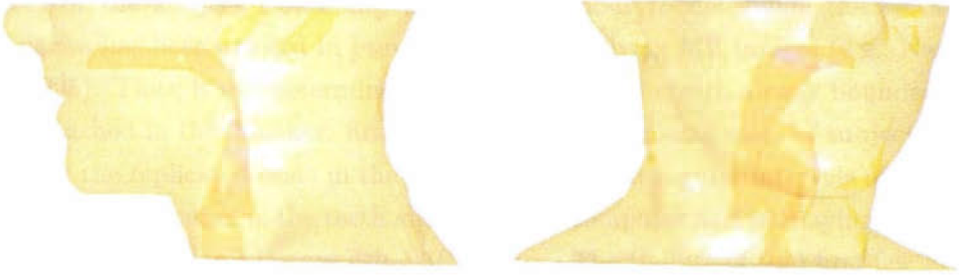

FIGURE 4.3: 3-D model of the head and vocal tract rendered from axial image slices. Lack of detail in the oral region is insignificant as the area measurements in that region are extracted from coronal slices.

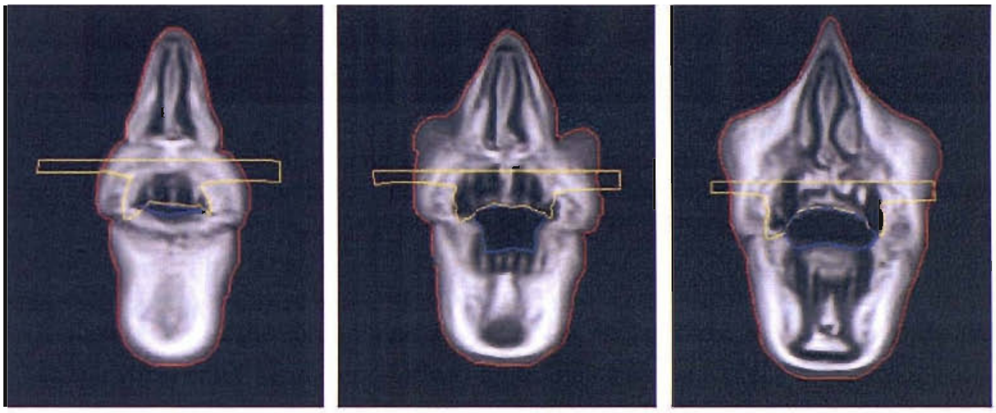

FIGURE 4.4: Determining the teeth-airway boundaries of the coronal slices. The boundaries of the upper incisors were determined using dental impressions of the subject. See text for more discussion.

exclusively from the coronal and axial models, respectively, while the area measurements at the vocal tract bend were derived by taking the mean of the areas from both models.

4.1.2 Determination of Teeth-Airway Boundaries

Chapter 2 explained how the dependence of the MRI technology on hydrogen molecules results in poor boundaries between areas with low concentrations of hydrogen, such as the teeth and the vocal tract airway. The teeth are surrounded by the vocal tract airway and careful consideration must be paid to differentiate between these areas. Figure 4.4 shows an example of three consecutive coronal slices from the oral region. The boundaries between the airway and the lower incisors have been determined, but one can see that they were not clearly defined.

The solution we employ utilises the dental impressions of the upper and lower incisors of the subject CS. Using dental impressions to determine the teeth-airway boundaries has been applied in previous studies involving MR images (e.g., Narayanan et al. 1995). Thus, it was determined beforehand that the teeth-airway boundaries must be approached in this manner: first, by replicating the dental casts of subject CS, then by slicing the replicated casts in the coronal direction at regular intervals of 4.0 mm per slice and finally, scanning the teeth slices into the computer and overlaying them on the corresponding (coronal) images using 3D-Doctor. The (failure of the) first few attempts to solve this problem drew several important points to our attention.

1. The material that is used to create the moulds must not shrink as it cures (i.e., hardens). In addition to this, it must be elastic enough to retain its shape, so as not to tear when taken off the original (plaster) impressions and when sliced. One highly used dental material called chromatic alginate, shrinks as it cures, tears easily when taken off the casts and suffers shape distortion when sliced. Thus, it was found to be unsuitable for our purposes.

2. Once the positive casts are made and sliced thinly, a single slice may result in several non-connecting fragments of teeth, specifically for the front-most region of the incisors. This requires additional work on their exact position for placement on the corresponding coronal image. Thus, a systematic method must be employed so that information about the exact position of each teeth segment is retained after slicing.

High-grade pourable silicone rubber was found to be the best material for this technique. It is a two-component silicone elastomer which cross-links at room temperature by poly-addition reaction. The silicone and the catalyst are viscous liquids, which, after mixing together and curing, become a strong elastic material. Its elasticity saves it from tearing when taken off the plaster casts and allows it to retain its shape when sliced. Most importantly, there is almost zero shrinkage during/after curing, ultimately producing a near exact replica of the original cast in a strong durable form.

To allow the teeth locations to be retained after slicing, the rubber impressions of the upper and lower incisors were appended to their respective moulds made from the same silicone rubber of a different colour to form a cube-shaped mould. The following describes the procedure more clearly:

1. Two small boxes with dimensions of 8.6 cm × 6.8 cm × 5 cm were made using laminated cardboard. Each box can fit neatly the original plaster casts of the upper and lower incisors of subject CS's dental impressions.

2. In a mixing container, the pourable silicone rubber was mixed with the catalyst. The mixture was poured equally into the two small (empty) boxes. Immediately

after pouring and well before setting (the mixture takes approximately 24 hours to cure), the plaster casts (of the upper and lower incisors) were placed teeth section down into the mixture in each box and left overnight to harden.

3. Once hardened, the original positive plaster casts were removed. The rubber impressions were left in the boxes.

4. In a mixing container, more pourable silicone rubber was mixed with the catalyst. This time black dry colour pigments were added to the mixture. (The colour pigments were dissolved in thinner first before being added to the mixture. This causes the mixture to be less viscous but does not affect its curing properties). The mixture was divided into two equal parts and poured *on top* of the rubber impressions in the boxes. It was then left to harden overnight.

5. The hardened mixtures form cubes that are half white and half black. As the box was laminated beforehand, the rubber cubes came off the walls of the box easily. They are placed in an electric Cookworks food slicer, which allows the user to set the thickness of the slices up to the nearest millimetre. The cubes were sliced to a thickness of 4.0 mm in the coronal direction. Each cube resulted in approximately 10 teeth slices.

The teeth slices were then scanned into the computer using a HP ScanJet. The images were processed in 3D-Doctor as JPEG images. The boundaries between the negative and positive casts were outlined. The negative cast surrounding the positive section allowed the location of the (front) teeth fragments to be retained after slicing. The process of matching the teeth boundaries to the correct slices was based on the distance of the image slice and teeth location from the lips and also the pattern of the tissue/gums surrounding the teeth in the images. An example of a teeth-boundary correction procedure on the upper incisors is shown in Figure 4.4. As can be seen in the figure, the square boundaries of the cube assist in aligning the teeth consistently from image to image.

Once the teeth as well as the face, neck and vocal tract boundaries had been determined, the coronal slices were rendered to form a 3-D model. An example of a completed 3-D model consisting of coronal image slices is shown in Figure 4.5 with the fixed teeth boundaries.

Before derivation of the area function, the coronal and axial models of the same fricative utterance were merged together in a single environment to form one complete head model for the fricative. This is why outlining the subject's face was an important part of the procedure; the ears, chin and nose act as location markers, which are used to align the models. A good match on the location of the ears, chin and nose confirms that the calibration parameters are correct and that the subject was in a consistent position while the (axial and coronal) images were acquired. The 3-D models were merged together for two important reasons:

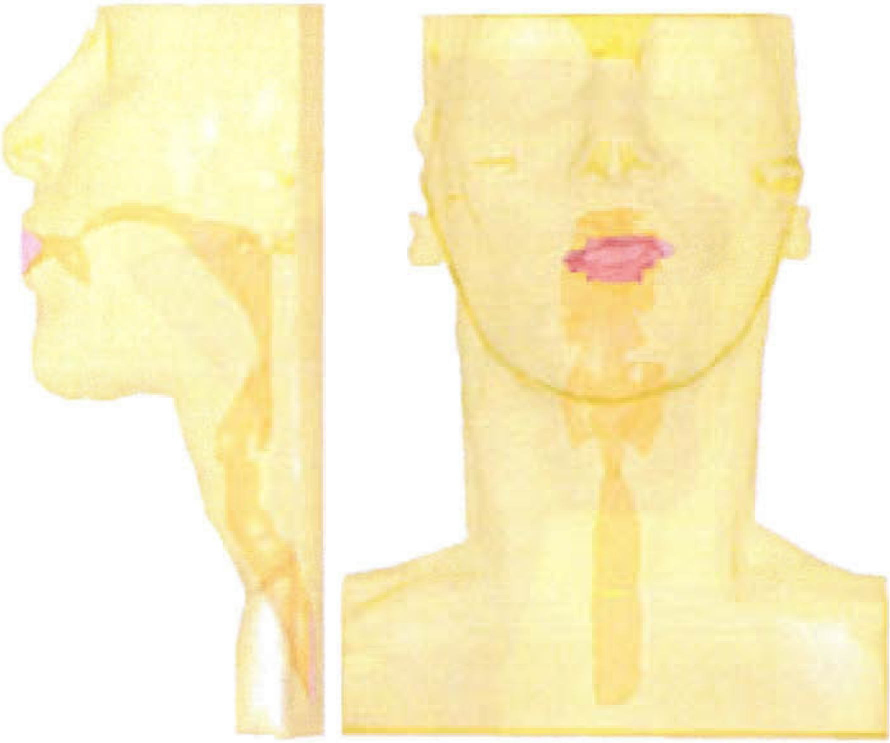

FIGURE 4.5: 3-D representation of the head and vocal tract derived from coronal image slices. Additional steps were taken to derive the teeth-airway boundaries because of the low hydrogen levels in these regions.

1. continuity in the location of the grid lines from anterior to posterior of tract; and
2. consistency in the location of the grid lines in the vicinity of the vocal tract bend, since the area measurements in this region are averaged.

An example, this time in wire frame, of a merged vocal tract and head model is shown in Figure 4.6 for subject CS while she sustained the fricative /θ/.

4.2 Implementation of Area-Derivation Techniques

The area-derivation techniques follow the two methodologies introduced by Mermelstein (1973) and Blum (1973), which have already been described in Chapter 2. The whole procedure involves the use of 3D-Doctor, CorelDraw and some computation by hand. In 3D-Doctor, a function known as "Cutting 3-D Contours" allows the placement of a plane anywhere in any angle within the 3-D environment and generates information on the area and perimeter of any object it encounters within that plane. This function was used extensively to derive the areas and perimeters of the vocal tract with the plane placed according to the locations and angles described by the area-derivation

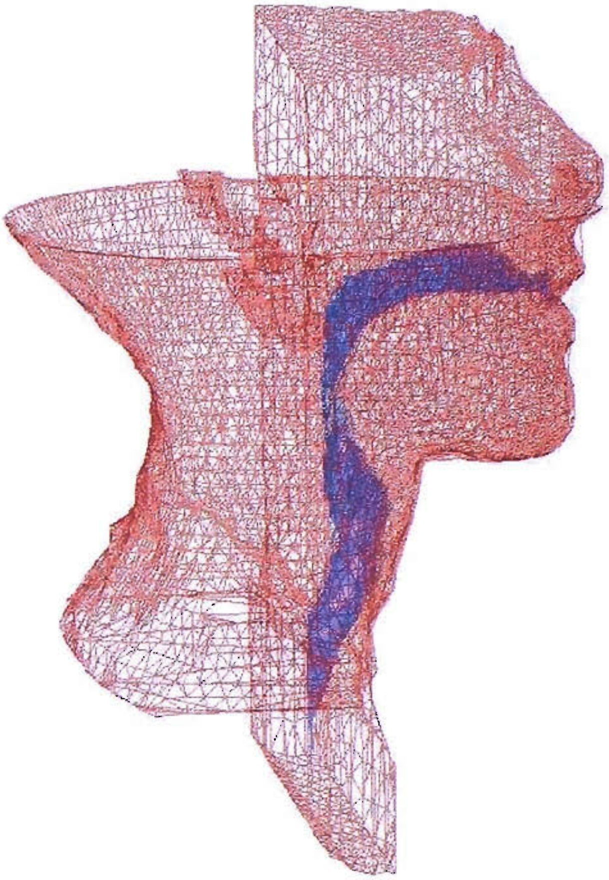

FIGURE 4.6: Merged 3-D models of head and vocal tract created from coronal and axial images. In this example, subject CS was sustaining the fricative /θ/.

methodology. An example of the application of this function in the 3-D environment is shown in Figure 4.7. In this figure, the blue lines on the 3-D model represent the grid lines positioned according to the Blum transform. The yellow plane is moved by the function to match the location and angle of the grid lines, and the measurements of any object it encounters within the plane are generated.

The following sections describe how the Mermelstein and Blum area-derivation techniques were implemented on the 3-D models. The terms *grid line* and *grid plane* will be used interchangeably and frequently. As previously mentioned in Chapter 2, each measured cross-sectional area is referred to as a grid plane but in a midsagittal diagram of the vocal tract, the grid planes are referred to as grid lines. In effect, the *grid lines* specify the location and angle of the *grid planes* in the midsagittal view.

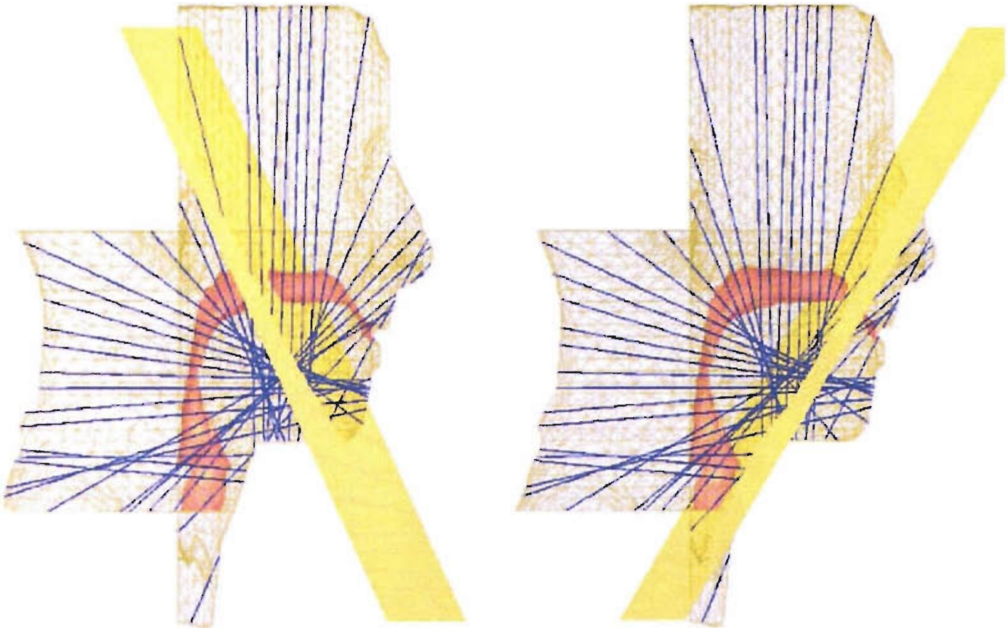

FIGURE 4.7: Application of "Cutting 3-D Contours" function. 3D-Doctor is able to derive area and perimeter information of all the objects on the specified plane within the 3-D environment by moving the location of the yellow plane. The blue lines represent grid lines placed according to the Blum area-derivation technique.

4.2.1 Mermelstein Technique

Remembering from Chapter 2 that the Mermelstein technique was based on the concept that the tongue is represented as a ball that has a fixed radius and a moving centre (Coker and Fujimura 1966). The necessary steps for this technique are as follows. A diameter for the tongue ball is identified that can represent all of the vocal tract configurations for the subject. Subsequently, in midsagittal, a circle representing the tongue ball was neatly placed into the curvature of the tongue boundary. The centre location of the tongue ball represents the centre point of the radial grid lines that are placed at intervals of 10° on the vocal tract bend, up to a full quarter of the circle. The vertical and horizontal grid lines of the vocal tract bend mark the beginning of the grid lines for the anterior and posterior region of the tract. These grid lines are parallel to each other and placed at 5.0 mm intervals. The midline of the vocal tract is now defined by connecting the centre point of each grid line. The difference in angle between each consecutive grid line is measured to obtain the cosine factor. The area measured at each grid plane is then multiplied with the cosine factor to obtain an estimated measurement of the area encountered by the propagating soundwave.

Implementation of the Mermelstein technique on the 3-D model is described as follows. To determine the diameter of the tongue ball, the 3-D models were rotated to provide a midsagittal view and hardcopies of the models in this angle were made for all the

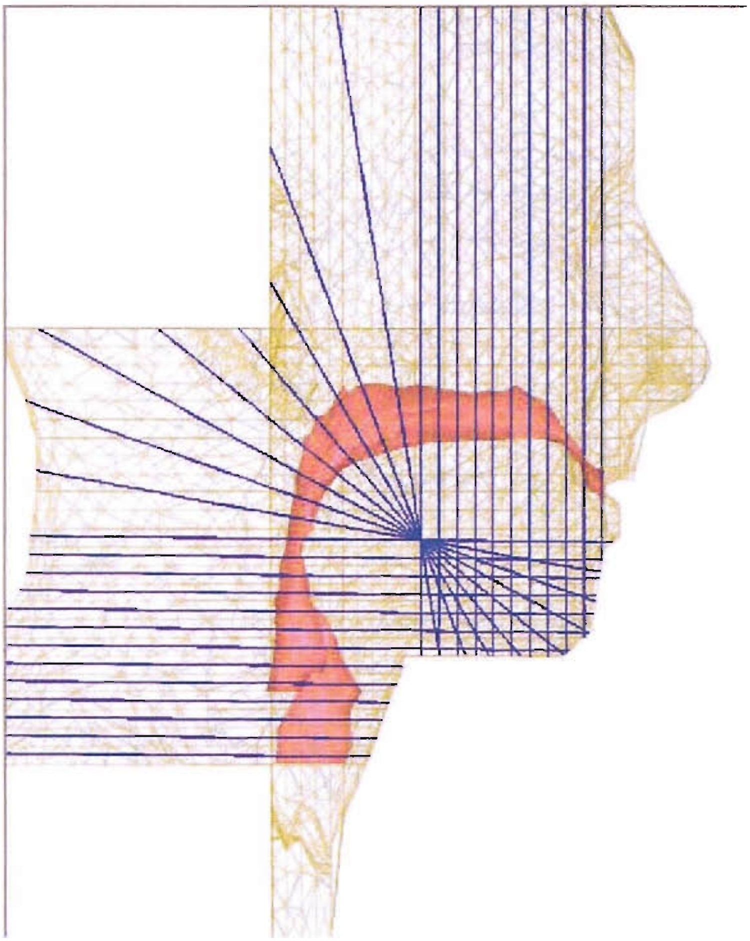

FIGURE 4.8: Application of the Mermelstein technique. Grid lines derived according to the Mermelstein technique. In this example, the subject was sustaining the fricative /s/.

fricative configurations. For each midsagittal model, a circle was fitted neatly against the shape of the tongue ball and the diameters were measured. The averaged diameter was computed, found to be 2.5 cm for subject CS.

Next, new circles were drawn (with a diameter of 2.5 cm) and fitted into the tongue shape of the midsagittal models to determine the centre point location of the radial planes. The coordinates of the centre point location were keyed into the 3D-Doctor environment, where the "Cutting 3-D Contours" function was used to generate the area measurements based on this centre point, by using radial planes at 10° intervals on the bend and vertical and horizontal planes, which are parallel for the anterior and posterior region of the tract. The built-in function was able to do this automatically (by specifying the slice intervals and location of the first slice) because the planes were in parallel or radial to each other. An example of the sliced 3-D model is shown in Figure 4.8.

Since the 3-D model consists of an axial-coronal pair, each measured grid plane will result in two types of measurement; one for the coronal model and one for the axial

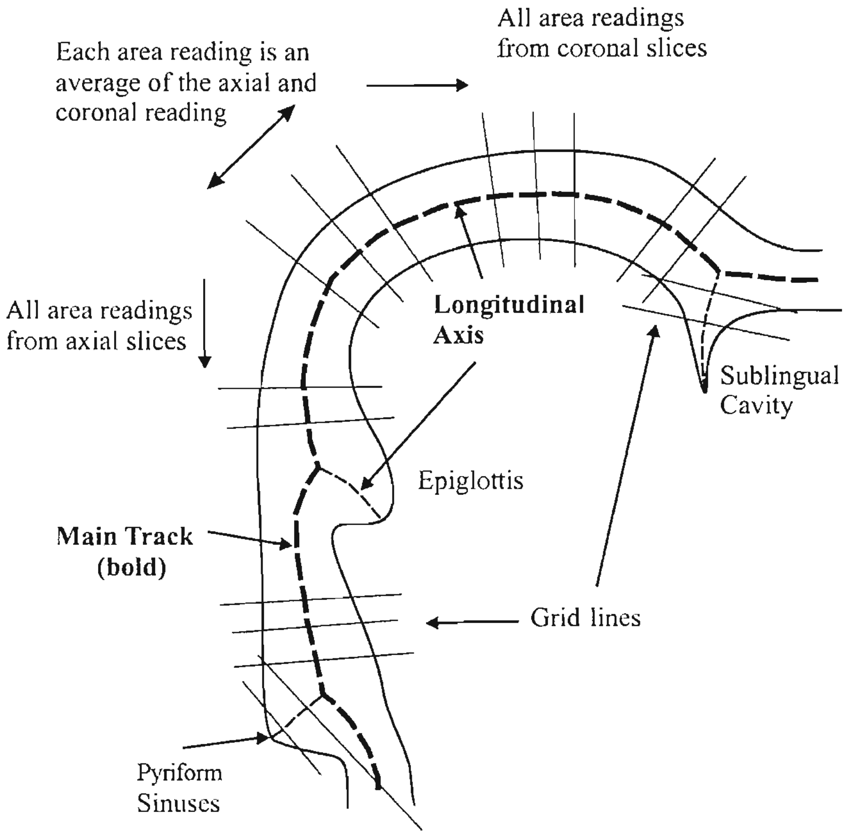
Each area reading is an average of the axial and coronal reading
All area readings from coronal slices
All area readings from axial slices
Longitudinal Axis
Sublingual Cavity
Epiglottis
Main Track (bold)
Grid lines
Pyriform Sinuses

FIGURE 4.9: Computation of the areas at the vocal tract bend, application of the Blum transform and definition of the terms used throughout this chapter. Assume the grid lines down the main tract are at 5.0 mm intervals. There are three branches in this example: the sublingual cavity, the epiglottis and the pyriform sinuses. The area measurements at the vocal tract bend are averages of the areas from the axial and coronal models. Anterior and posterior to this bend, the area measurements were derived from the coronal and axial models respectively. Note the locations of grid planes at junctions. In some cases, two grid planes may cover the same area.

model. For the derivation of the area function in this investigation, the areas from the lips to the pharynx were taken from the coronal model, while areas from the pharynx to the glottis were taken from the axial model. At approximately 130° to 150° of the vocal tract bend, refer to Figure 4.9 for clarity, the cross-sectional area of each grid plane was determined by averaging the areas of the coronal and axial models represented by $A_j = \frac{A_j^c + A_j^a}{2}$, where A_j^c and A_j^a are the cross-sectional areas of the jth slice of the coronal and axial slice respectively.

Once the area measurements were derived, the 3-D models were rotated to midsagittal and new hardcopies of the model in this position were made. The midline is traced along the length of the vocal tract by connecting the centre point of each grid line. The midline may be similar to the longitudinal axis but it serves a different purpose; between two consecutive grid planes, the difference between the normal of the first grid plane and the angle of the midline is measured to obtain the cosine factor. Then, each cross-sectional area measurement is multiplied to the corresponding cosine factor to give an estimate of the area encountered by the soundwave as it travels through the tract.

The Mermelstein technique treats the side branches differently from the Blum technique. Unless the areas of the side branches are connected to the main airway on the measured grid plane, they are completely ignored. An example of this can be seen from the outlined image shown previously in Figure 4.2. For this particular grid plane, the Mermelstein-derived area includes the main tract and the side branch labelled section A because it is connected to the main tract. The area of section B is not included in the area function.

4.2.2 Blum Transform

The necessary steps for this transform are as follows. First, the derivation of the main longitudinal axis of the vocal tract as well as the longitudinal axis of the side branches must be done using circles placed throughout the vocal tract borders while in midsagittal view. At least two points on the circle must be tangent to the walls of the vocal tract. Next, the longitudinal axis is derived by connecting the centre points of the circles. Finally, grid lines perpendicular to the longitudinal axes are placed at intervals of 5.0 mm, and area measurements are taken at the location of these grid lines.

Figure 4.9, shown previously, illustrates the technique and terminology related to this slicing procedure. The location of the grid lines and the side branches that are shown in this figure are for explanatory purposes only. The side branches were emphasised because their areas were also derived. However, unlike this example, the data processed involved a maximum of two side branches: the pyriform sinuses and the sublingual cavity.

Implementation of the Blum area-derivation technique on the 3-D models is as follows. First, a merged 3-D model of an axial-coronal pair was rotated on the $y-$axis so that the model was viewed from midsagittal. A snapshot of the model at this angle was made and the figure was exported into the CorelDraw environment. Using CorelDraw, pre-drawn circles with distinct centre points were manipulated to fit into the vocal tract boundaries, with at least two points on the circle tangent to the vocal tract boundaries. The circles were placed closely together and by connecting the centre points of the circles, the longitudinal axis was traced from the lips to the glottis. Still within the CorelDraw environment, grid lines with intervals of 5.0 mm were placed perpendicularly along the longitudinal axis. For clarity, an example of this procedure is shown in Figure 4.10. In (a), a midsagittal model was fitted with circles and the longitudinal axis was derived. In (b), the derived data consisting of the longitudinal axis and the grid lines are shown. The longitudinal axis and the grid lines shown in Figure 4.10(b) were extracted from the CorelDraw environment and exported back to the 3D-Doctor environment. The

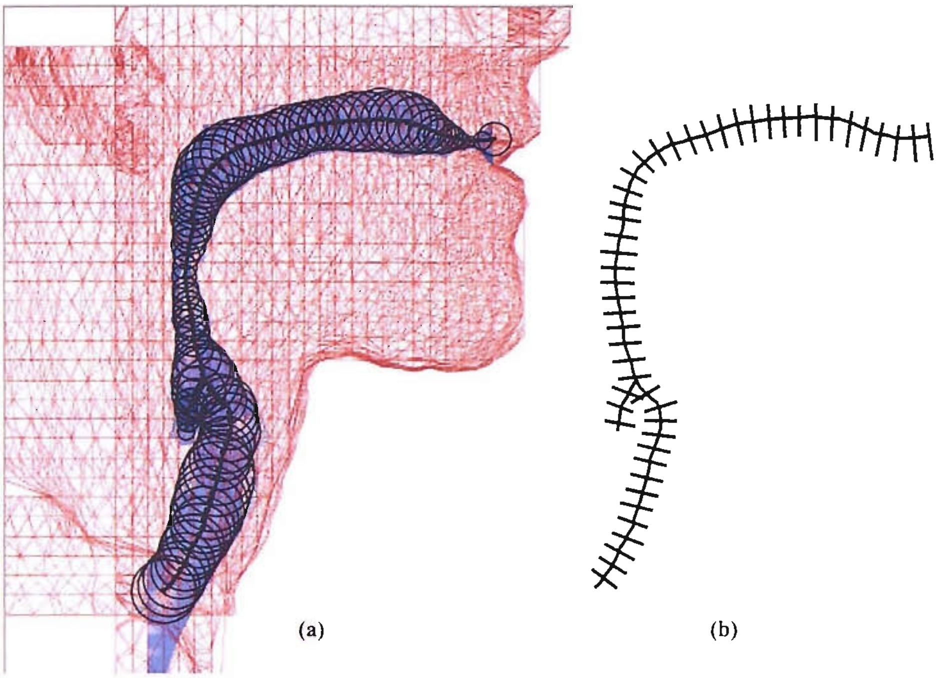

FIGURE 4.10: (a) Application of the Blum transform on the 3-D model which was done in midsagittal view. The circles were placed using the CorelDraw drawing software. (b) The longitudinal axis of the vocal tract was derived from the centre points of the circles. Subsequently, the grid lines were placed along the midline at angles normal to the plane. The longitudinal axis and the location of the grid lines were extracted from the CorelDraw environment and exported to the 3D-Doctor environment to derive the area function.

"Cutting 3-D Contours" function was used to extract the area measurements and perimeters at the locations and angles specified by the grid lines.

Since the 3-D model consists of an axial-coronal pair, each measured grid plane will give two types of measurements: one for the coronal model and one for the axial model. For the derivation of the area function in this investigation, the areas from the lips to the pharynx were taken from the coronal model, while areas from the pharynx to the glottis were taken from the axial model. At approximately 130° to 150° of the vocal tract bend, again refer to Figure 4.9 for clarity, the cross-sectional area of each grid plane was determined by averaging the areas of the coronal and axial models represented by $A_j = \frac{A_j^c + A_j^a}{2}$, where A_j^c and A_j^a are the cross-sectional areas of the jth slice of the coronal and axial slice respectively.

As can be seen from Figure 4.9, areas of the side branches are derived using a separate longitudinal axis and because of this, they will be presented as separate areas in the figures (of the results section), with regard to distance from the lips. In the region of the pyriform sinuses, a single grid plane may give three different measurements of the axial model: the main area of the vocal tract and the two areas of the pyriform sinuses. In this case, the areas of the side branches are consistently added together to form one area of the side branch. Sometimes however, the area of the side branch is connected to the main airway on that particular plane. An example of such an occurrence can be seen in Figure 4.2 (shown earlier in this chapter). This example includes the connected area (section A) as part of the main tract and treats the disconnected area (section B) as a side branch.

4.3 Hydraulic Radius

The hydraulic radius parameter, R_H, is the degree of *roundness* of a cross-sectional area. The reason for calculating the R_H parameters is to compute the losses of each vocal tract segment caused by surface absorption. If the cross-sectional shape was completely round, the R_H radius would approach unity, and thus minimal losses would occur from surface absorption (Davies, McGowan, and Shadle 1993). However, as can be seen in Figure 4.1, most sections of the vocal tract have non-circular cross-sectional areas. Thus, when ACTRA calculates the transfer functions from the area functions, the hydraulic radius parameters are used to estimate the surface losses of each segment within the vocal tract.

The hydraulic radius parameter is computed using the area and the perimeter measurements of the vocal tract. It is given by:

$$R_j = \frac{2A_j}{P_j}, \tag{4.1}$$

where A_j and P_j are the area and the perimeter of the jth grid plane respectively. Therefore, each cross-sectional area of the area function will have a corresponding hydraulic radius parameter.

4.4 Results and Discussion: Area Functions

The area functions derived using the Blum transform and the Mermelstein technique are presented in Figures 4.11 to 4.16 for the fricatives /s, ʃ, θ, (a)f, (i)f/ and /(u)f/ respectively. The figure for each fricative contains three plots, namely: (a) a bar plot of the area derived through the Mermelstein technique; (b) a bar plot of the area derived using the Blum transform; and (c) a bar plot of the areas of side branches derived using the Blum transform.

In a typical vocal tract geometry, the pyriform sinuses bifurcate into two cone-shaped columns off the laryngeal wall, but in the subplots labelled (c) presented in each figure, the areas of the pyriform sinuses (derived from a single grid plane) were summed

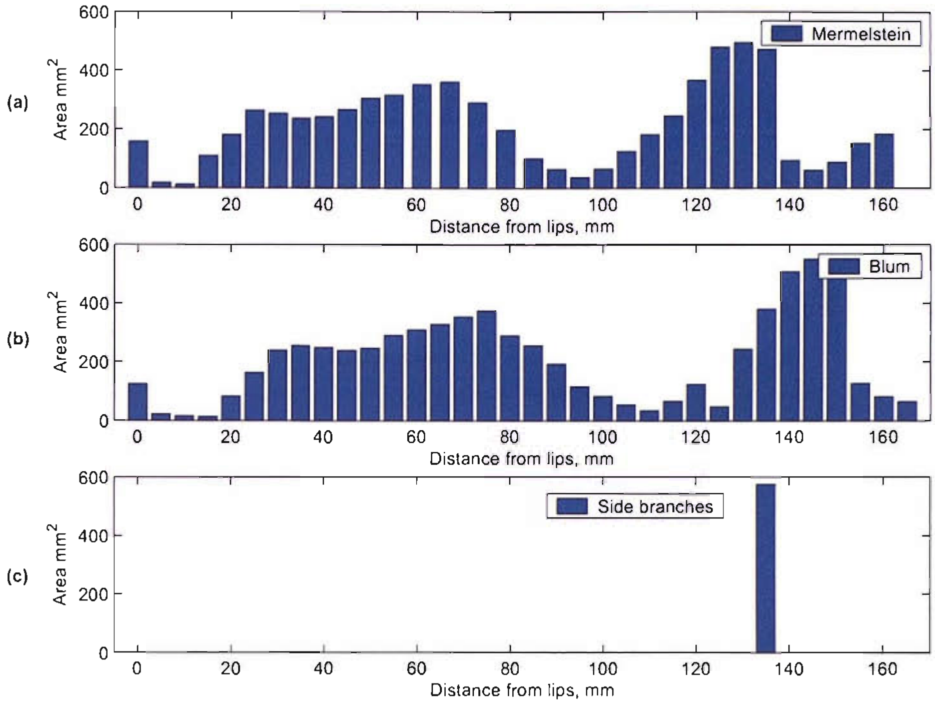

FIGURE 4.11: Area functions of /s/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

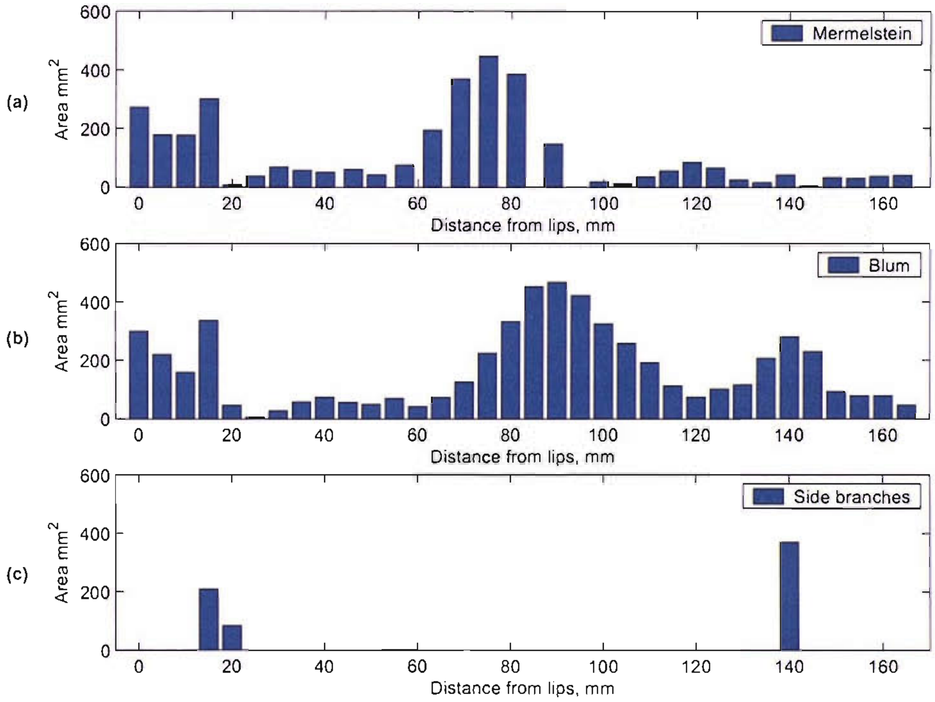

FIGURE 4.12: Area functions of /ʃ/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

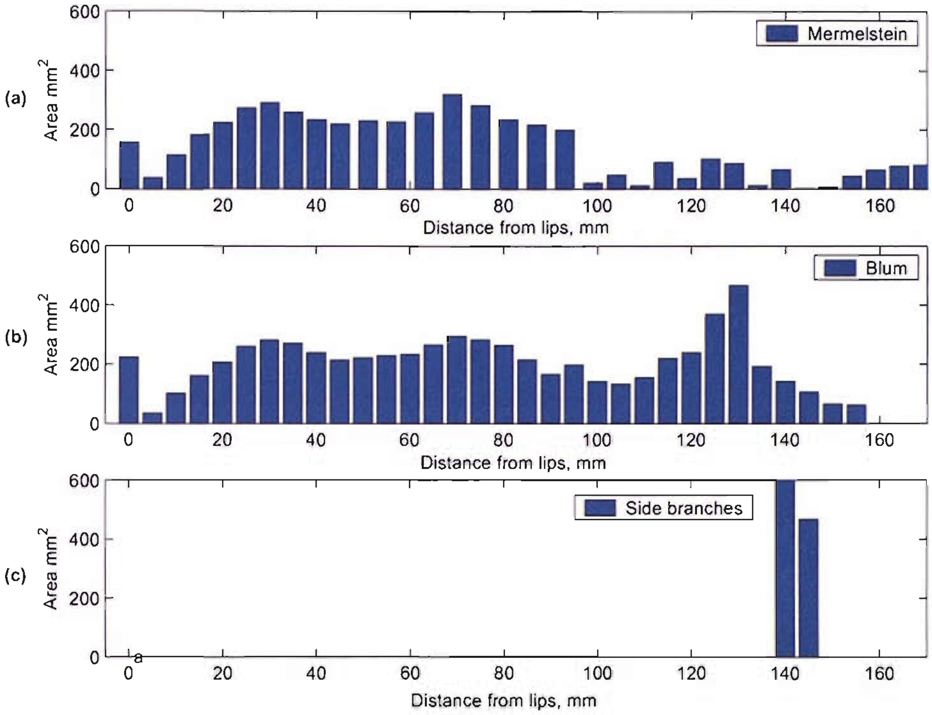

FIGURE 4.13: Area functions of /θ/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

together. This is because they will be modelled as one area for calculation of the transfer functions.

The area functions are a function of distance from the lips to the glottis. Therefore, areas of side branches within the range of 0 to 40 mm represent the sublingual cavity and within the range of 120 to 160 mm represent the pyriform sinuses in the laryngeal region.

4.4.1 Mermelstein-derived Area versus Blum-derived Area

The main difference between the Mermelstein- and Blum-derived area functions is the areas of the side branches for Blum, which are presented separately from the areas of the main tract. The Mermelstein technique ignores side branches unless they are connected to the main airway within the measured plane. However, two other major differences were noted between the Mermelstein and Blum-derived area functions:

1. The different approaches that were taken to position the grid lines result in more readings for the Blum transform and less for the Mermelstein technique. This is because the Blum transform places grid lines at 5.0 mm intervals along a longitudinal axis which is irregular, since it follows the (midsagittal) shape of the

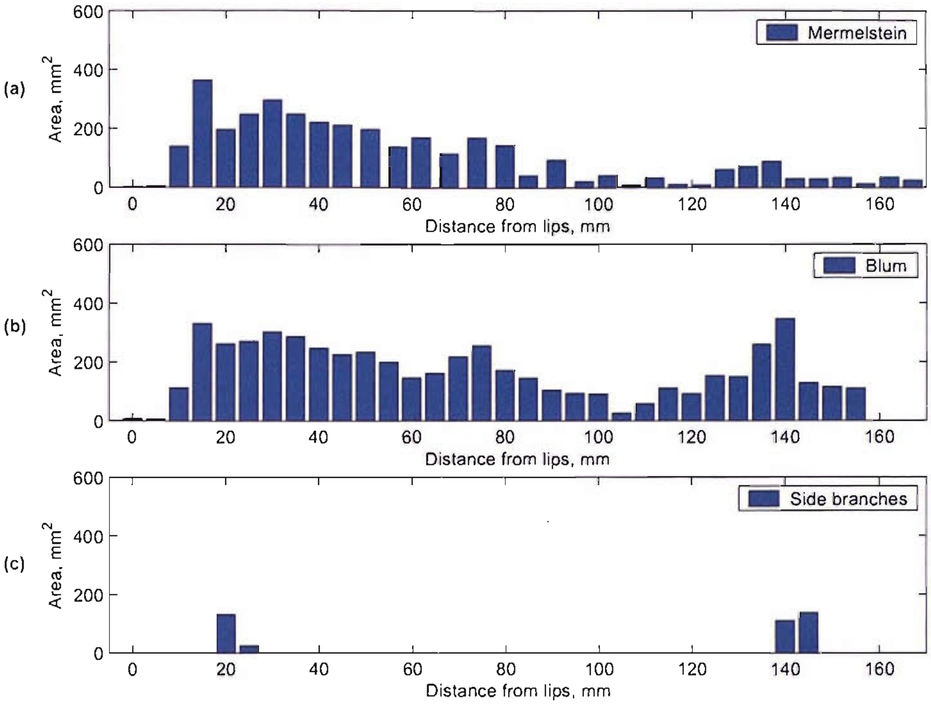

FIGURE 4.14: Area functions of /(a)f/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

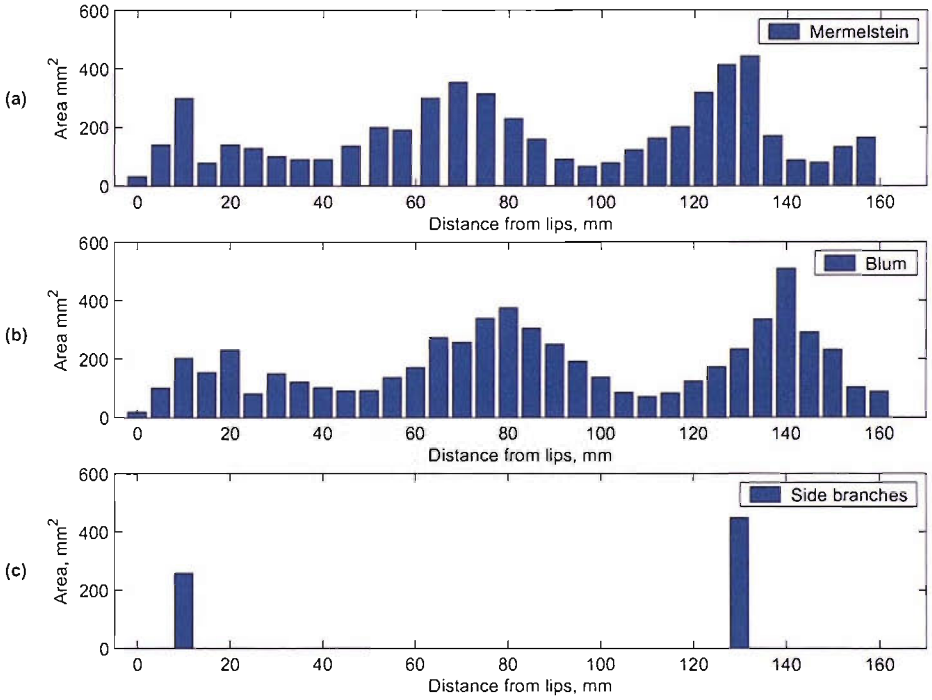

FIGURE 4.15: Area functions of /(i)f/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

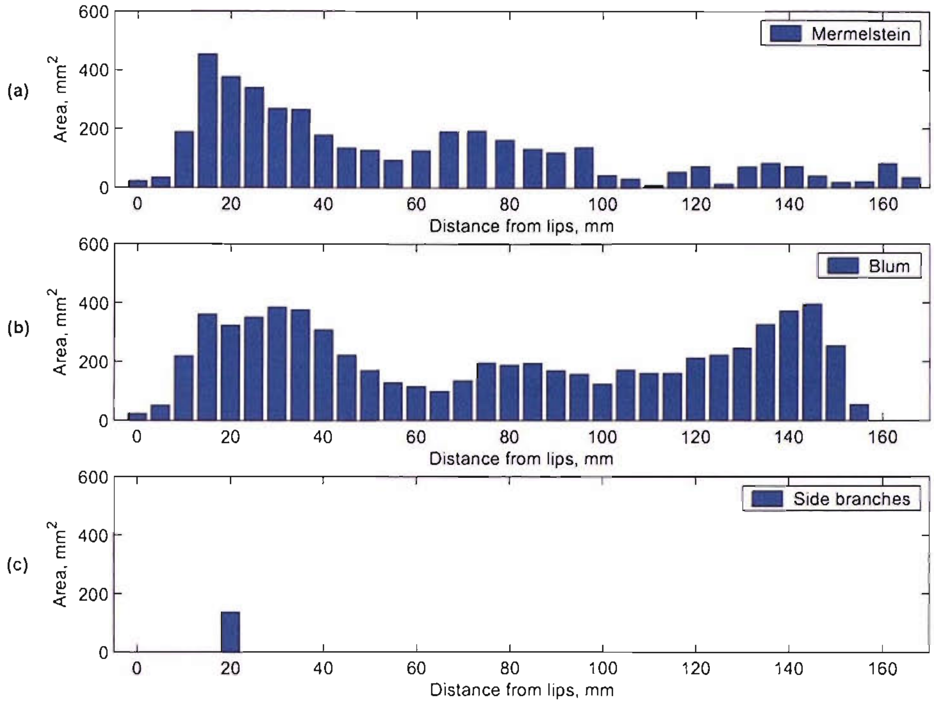

FIGURE 4.16: Area functions of /(u)f/. (a) Area function derived using the Mermelstein technique. (b) Area function derived using the Blum transform. (c) The areas of the side branches derived using the Blum transform.

tract. The Mermelstein technique does not derive a longitudinal axis but places the grid lines at 5.0 mm/10° intervals parallel/radial to each other (depending on the region).

2. Some of the Mermelstein-derived area functions show substantially lower area measurements, particularly in the posterior region of the vocal tract compared to the Blum-derived ones. This can be seen in the subplots labelled (a) for the fricatives /ʃ, θ, (a)f, (u)f/ (Figures 4.12, 4.13, 4.14 and 4.16, respectively), whereas higher area measurements are observable for /s, (i)f/ (Figures 4.11 and 4.15, respectively). The probable cause for this difference is the difference in the angles with which each grid line is placed between the two area-derivation techniques.

Apart from the differences noted above, it was also noted that, depending on the placement of the tongue ball, the radial segments at the vocal tract bend of the Mermelstein area-derivation technique may not necessarily create 5.0 mm intervals on the midline. Therefore, the plots of the Mermelstein-derived areas were altered to reflect the correct distance along the vocal tract bend.

Finally, it was noted that applying the Mermelstein technique requires less work than the Blum technique; the regular intervals between the parallel planes simplify the implementation of the technique in 3D-Doctor and make the procedure less prone to errors (since the grid planes were placed automatically by 3D-Doctor). For the Blum technique, the placement of circles for each vocal tract configuration requires extensive work and manual marking out of the location of perpendicular grid lines at intervals of 5.0 mm on an irregular axis is prone to errors.

4.4.2 Sibilant Fricatives

Figure 4.11 shows the area functions of /s/. The area functions show that the constriction is about 10–15 mm long. This is expected, since the tongue tip is elevated and rests against the roof of the mouth to form the constriction. Shadle (1985) noted variations of the /s/ configuration across utterances for a single speaker. The tongue tip location can be either behind the upper incisors or the lower incisors. In this case, the lack of sublingual cavity suggests that the subject had placed her tongue tip behind the lower incisors to form the constriction.

In general, the area functions from the two techniques show similarity in the area measurements of the main tract. Differences seem to lie in the distances with regard to the lips. The Blum-derived area has a longer constriction length than the Mermelstein-derived area, as with the overall length of the tract. The cause for this was noted in the previous section.

The area functions of /ʃ/ are shown in Figure 4.12. This fricative has the longest constriction, approximately 40 mm altogether, with the smallest area occurring in the anterior part of the constriction region. The production of /ʃ/ requires the tongue blade up to the dorsum to elevate and rest against the lateral dome of the hard palate for the formation of a long narrow channel for the constriction. This is reflected in the area functions. The cross-sectional area posterior to this channel, near the tongue root and adjacent to the vocal tract bend, is relatively wide compared to other fricative configurations. The Blum-derived area function shows the existence of a sublingual cavity. This is expected of the /ʃ/ configuration, because of the elevated tongue tip.

4.4.3 Nonsibilant Fricatives

The area functions of /θ/ are shown in Figure 4.13. As expected, the smallest constriction was seen to occur in the anterior region of the oral cavity, where the tongue is placed between the upper and lower incisors to form the constriction. The length of the constriction is expected to be short, shown to be 5.0 mm in the plots, but the area posterior to the constriction is also very small, as the tongue body was slightly elevated. No sublingual cavity was formed during production. The Mermelstein- and Blum-derived areas show similarities in the area measurements for the anterior region of the tract up until after the vocal tract bend.

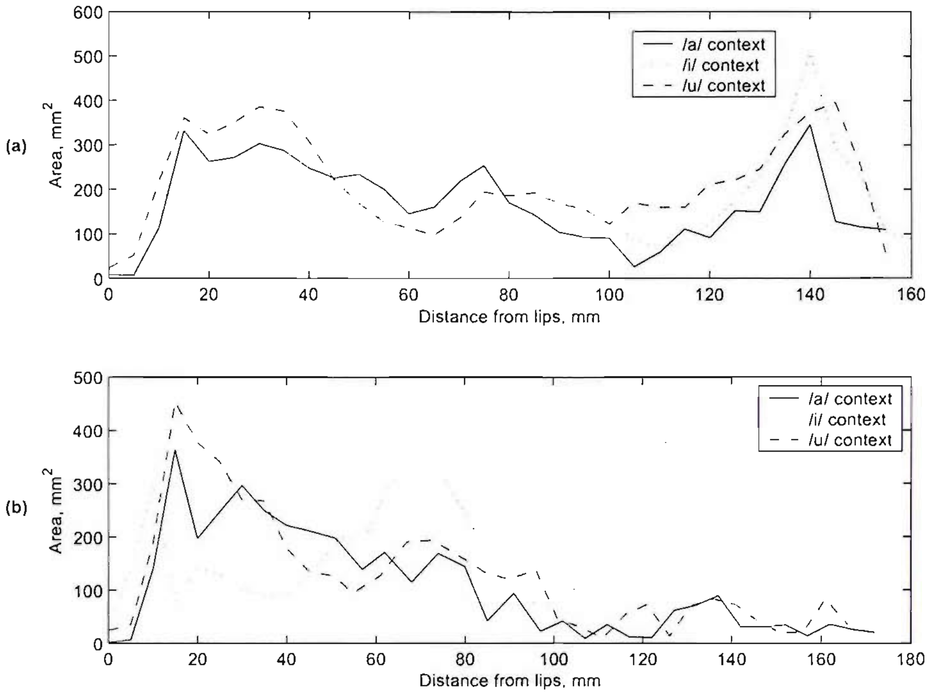

FIGURE 4.17: Area functions of /f/ in different vowel contexts derived according to: (a) the Blum transform; and (b) the Mermelstein technique.

The area functions of /f/ are shown in Figures 4.14–4.16 for the vowel contexts /a, i, u/ respectively. For the labiodental fricative, the smallest point of constriction occurs in the front-most region of the oral cavity, where the upper incisors rest against the lower lips. Thus, the area measurements commence approximately 5.0 to 10 mm (depending on context) anterior to the upper incisors.

In all productions of /f/, subject CS was observed to form a sublingual cavity between the tongue and the lower incisors. The Blum measurements show that the cavity area measures 140, 250 and 140 mm^2 for the vowel contexts /a, i, u/ respectively. In the Mermelstein-derived area functions, it is highly probable that the higher area measurements in the oral region represent the area of the sublingual cavity. This can be seen in the 15–20 mm bars in the subplots labelled (a) in the /f/ figures which give a measurement of approximately 350–390 mm^2, implying that the sublingual cavities may be approximately the same size.

To look at the differences in the area functions of /f/ in the three vowel contexts, the area functions were overlaid in one plot for each area-derivation technique, shown in Figure 4.17. Subplot (a) shows the Blum-derived area functions and subplot (b) shows the Mermelstein-derived area functions.

It is evident that the area measurements of /f/ in the three vowel contexts are different throughout the tract. It can be said quite confidently that this is the effect of vowel context on the fricative, because with the exception of the constriction in the lip region, the area functions have similar characteristics to those of their respective vowels (see Baer et al. 1991, for area functions of vowels). For example, for vowel context /i/, which is classified as a "high-front" vowel, the tongue is thrust upward to form a narrow constriction which effectively reduces the area in the oral region and increases the region in the pharyngeal area. For /u/, a "high-back" vowel, the back of the tongue body is elevated resulting in a larger area in the oral region and a wide pharyngeal cavity with some constriction at the velar. For /a/, a "low-back" vowel, the mouth is opened and the tongue body lowered, resulting in a larger front cavity and a narrowing of the pharynx (Baer et al. 1991; Stevens 1998).

The features described for each vowel can be observed in the corresponding area measurements of both area-derivation techniques. These are:

1. the smaller area in the oral cavity for /i/ context; the larger areas for the /a, u/ contexts; and
2. the larger area in the vicinity of the pharynx for /i/, the smaller areas for the /a, u/ contexts.

Clearly, the area measurements of the Mermelstein and Blum techniques are different. However, both managed to capture features that are similar to that of the vowel in which the fricative was uttered. The obvious differences between the area functions derived using the two techniques lie in the vicinity of the pyriform sinuses. The Blum transform derived similar measurements in this region, whereas the Mermelstein technique measured lower values for the /a, u/ contexts. As previously noted in Section 4.4.1, the cause for this can only be attributed to the area-derivation technique, because the areas for each fricative were derived from the same models. It was noted that other investigations using the Mermelstein technique had slightly angled the vertical/horizontal planes to better match the angle of the vocal tract when deriving the area functions (e.g., Badin et al. 1998).

Comparisons of the area functions derived here were made with those of Narayanan et al. (1995). The method that was used to derive the area functions in their investigation is similar to the Mermelstein technique, as the area measurements were taken directly off the coronal and axial slices for the anterior and posterior region of the vocal tract respectively. For the anterior region starting from the lips up to the vocal tract bend, the area functions of Narayanan et al. (1995) show similarities in gross features to the area functions of both the Mermelstein and Blum-derived areas, but for areas after the bend, they matched those of the Blum-derived area functions better, that is with higher area measurements in the vicinity of the pyriform sinuses.

If the simplicity of the Mermelstein technique—brought about by the assumption that the bend is exactly 90°—produces less accurate area measurements, then this does not imply that the complexity of the Blum transform produces more accurate results. Because the grid planes are always placed perpendicular to the longitudinal axis, in regions where the axis is abruptly curved—such as the pyriform sinuses—it was observed that in some cases the same area is measured twice because of an overlap in the grid planes. (Refer to Figure 4.9 for an example, in which, note the grid lines that were placed on the longitudinal axis of the pyriform sinuses and the sublingual cavity. The grid planes closer to the main tract will encompass areas that have already been included as part of the main tract.)

4.4.4 Sublingual Cavity and Pyriform Sinuses

For subject CS, the formation of a sublingual cavity depends on the fricative uttered. For /ʃ/ and /f/ in all vowel contexts, a sublingual cavity formed during production. For /θ/ and /s/ a sublingual cavity did not form during production. The sublingual cavity for /ʃ/ is expected, because the tongue blade is raised. For fricatives /f/, /θ/ and /s/, the existence and size of a sublingual cavity is speaker dependent (Narayanan et al. 1995; Shadle et al. 1996).

The areas of the pyriform sinuses for all the fricatives were derived, except for /(u)f/, because a separate longitudinal axis for the pyriform sinuses could not be defined from a midsagittal view. In this case, the pyriform sinuses were located on either side of the vocal tract and could only be distinguished if the model was slightly rotated on the y−axis as shown in Figure 4.18.

The pyriform sinuses show very different area measurements for each fricative. There may be two reasons for this. The first and most likely reason is that pyriform sinuses have variable sizes which alter during articulation (e.g., Baer et al. 1991; Narayanan et al. 1995). The second reason is related to the implementation of the Blum area-derivation technique. Because of the cone-like shape of the pyriform sinuses, if the grid planes of the side branch were located higher up on the structure, the area measurement would be larger and may also encompass some parts of the main tract. If they were placed halfway down the middle of the structure, the area would be lower. But since the grid planes were placed at fixed intervals along the length of the tract, the inconsistencies in the location of the grid planes across different fricative geometries could not be avoided.

4.5 Summary and Conclusions

This chapter described how the Mermelstein and Blum area-derivation techniques were implemented on the 3-D images. The process requires the use of 3D-Doctor, CorelDraw and manual computation. The boundaries of the subject's profile and vocal tract airway were outlined from coronal and axial volumetric images. The teeth-airway boundaries

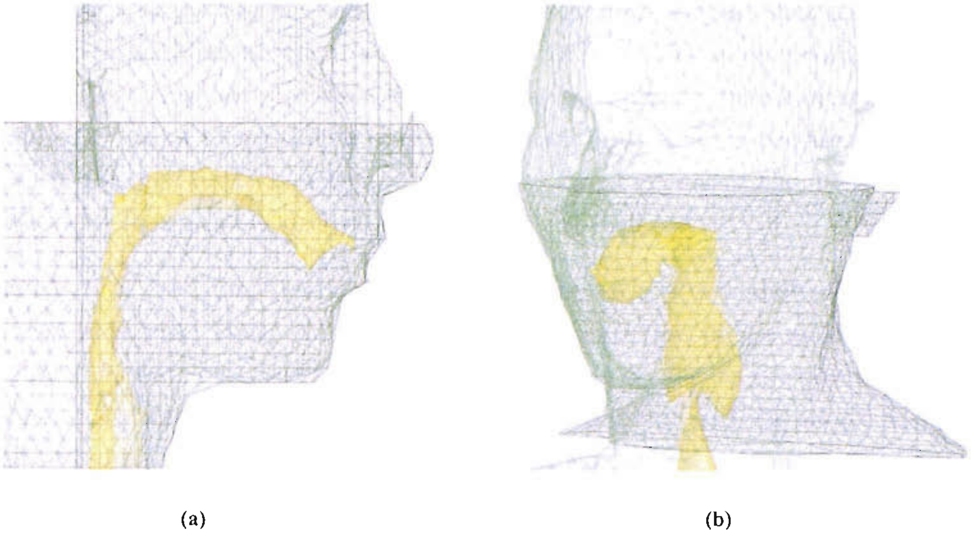

FIGURE 4.18: There were problems deriving the areas of the pyriform sinuses of /(u)f/. (a) In the midsagittal view the pyriform sinuses are located on either side of the vocal tract and cannot be distinguished from the main tract in midsagittal view. (b) They can only be seen when the 3-D model is rotated to provide a three-quarter sagittal view. For /(u)f/, failure to derive a separate longitudinal axis for the pyriform sinuses means that the areas were not derived.

were determined using positive and negative casts of the subject's dental impressions. The casts (upper and lower incisors), in the form of cubes, were subsequently sliced and scanned into the computer. In 3D-Doctor, the teeth boundaries were matched to the corresponding image slice where the teeth boundaries were unclear. Once the boundaries of the vocal tract airway were completed, the coronal and axial images were rendered and merged together to form a 3-D model of the subject's head and vocal tract.

The area-derivation techniques were implemented on the 3-D models using the built-in function "Cutting 3-D Contours". The location of the grid lines on the 3-D model was determined from a midsagittal view. The Blum technique was highly dependent on the longitudinal axis for the placement of the grid lines, while the Mermelstein technique depended on the midline to compensate for the slight curvatures of the vocal tract. It was evident that the two area-derivation techniques produced area functions that were different from each other, albeit derived from the same 3-D model.

Comparison of the area functions obtained from the Mermelstein and Blum area-derivation techniques has shown that the two major differences between them lie in the area measurements of the laryngeal region and the computed distance of the area function. The methodology of the Blum technique creates grid planes based on the longitudinal axis of the vocal tract, while the Mermelstein technique locates the grid planes first before deriving the midline. This will result in more grid planes for the Blum technique, because of the irregularity and/or bend of the vocal tract axis and fewer area measurements for the Mermelstein technique. The inconsistencies in the laryngeal measurements of the Mermelstein-derived area functions suggest that the subject may be in a slightly different posture between the different fricative recordings and that the cosine factor has not provided adequate compensation for the curvature in the tract. This does not mean to say that it is an incorrect representation of the area; it remains to be seen which area function will produce spectra that best match the measured spectra, the Blum- or Mermelstein-derived area.

Comparison of area functions for /f/ in different vowel contexts show differences in the area measurements. With the exception of the constriction in the anterior region, the area functions were observed to have features that are similar to the vowel in which the fricative was uttered. This supports the findings of Subsection 3.7.2, where the fricative was observed to be most influenced by vowel context.

The Blum transform is unique because its methodology includes the derivation of the areas of side branches. This will allow us to model the branches realistically for the derivation of the predicted spectra. The Mermelstein technique includes areas of side branches only if they are connected to the main tract on the measured grid plane. The advantage of this technique is that it is simpler to implement and thus less prone to errors. However, the Blum transform was also highly simplified because it was adapted according to Goldstein (1980), who worked only with midsagittal images. More accurate measurements could be obtained if the longitudinal axis of the pyriform sinuses was derived with regard to a coronal view. However, even if this were accomplished, the derived information cannot be used to its full potential, because the acoustic prediction software will model the side branches as a single area. Highly accurate representations of the back cavities may not be necessary if the cavities are decoupled during production.

The following chapter introduces the acoustic prediction software, ACTRA, used to calculate the transfer functions from the area functions. In short, by taking the combined product of the transfer function with a radiation characteristic term and a source model, we obtain the predicted spectrum of the fricative.

Chapter 5

ACTRA: Vocal Tract Acoustics Program

The previous chapter described the implementation of the Mermelstein and Blum area-derivation techniques on a 3-D image corpus and subsequently derived the area functions. The next step is to calculate the transfer functions from the area functions, compute the predicted spectra by taking the combined product of the transfer function (TF) with a radiation characteristic term and a source spectrum, and compare the predicted spectra to the measured spectra of Chapter 3. This chapter describes the acoustic prediction software used to calculate the TFs. It is known as ACTRA.

ACTRA was originally developed by Davies (1988) as a tool to model the acoustic behaviour of an automotive exhaust system. It was later adapted by Davies et al. (1993) to model vocal tract acoustics by processing area functions derived from an MR image corpus of vowel utterances. Jackson (2000) used ACTRA on a fricative corpus; the area functions were derived from midsagittal images, the perimeter of the area being transformed to represent the volume of the vocal tract airway. However, unlike this study, Jackson's investigation was based on the characterisation of plosive, fricative and aspiration sounds, while this work focuses on the perceptually important features of the fricative spectra and articulatory properties.

This chapter describes the acoustic assumptions that the software is based on, and the parameters that are defined in the program which enable it to simulate the vocal tract environment. A description is given of how a pressure source is used to represent the noise source in a fricative model and an example provided of how the vocal tract transfer functions (VTTFs) of the /s/ area function are calculated. It is hoped that from this chapter, the reader will gain a better understanding of the fricative spectra and the derivation of the synthetic fricative sound.

Because this in an introductory chapter on the software, the calculation of the TFs will be limited to a single pressure source location. The following chapter describes how the predicted fricative spectra are modelled by the superposition of several transfer functions of varying pressure source locations combined with a source model. This effectively models a distributed source spectrum that is more accurate to fricative production than a single concentrated source. This is because the frication noise is distributed downstream of the constriction and not concentrated on a single location.

Appendix A documents some preliminary tests that were done on ACTRA using several models from the literature where the acoustic response of the models have been measured or can be computed and subsequently, comparisons are made with the the transfer functions that are calculated by ACTRA. These models range from simple uniform tube configurations to a complete area function taken from Narayanan and Alwan (2000). The results of these tests enable use of the acoustic prediction software with confidence.

5.1 Introduction to ACTRA

There are several reasons for choosing ACTRA to calculate the TFs for this investigation:

1. The program allows the user to define an intermediate pressure source anywhere within the vocal tract model. This is essential for fricative modelling; the pressure source creates a closed-end effect, allowing calculation of the system zeros. The role of the pressure source will be described in more detail in the following sections.

2. The program is able to incorporate side branches in a relatively realistic manner (as opposed to an abrupt increase in the area). The areas are modelled as an orifice coupled to the main tract at one end and closed at the opposite end. This allows investigation on the importance of realistically modelling the side branches for fricative synthesis.

3. The program is able to incorporate hydraulic radius parameters. As previously explained in Section 4.3, these parameters are used to estimate the losses that occur as a result of surface absorption at regions where the cross-sectional areas of the vocal tract are non-circular.

In ACTRA, the vocal tract area function is modelled as a concatenation of (predefined) elements whose length and area measurements can be altered by the user to match each section of the vocal tract. The program derives the VTTFs on the assumption of 1-D planar wave propagation. The curve of the vocal tract is ignored, because the radius of curvature at the bend is sufficiently large compared to the axial dimensions of the tract and the effects are considered negligible. Three assumptions of classical acoustic speech theory applied to the vocal tract are relaxed by ACTRA:

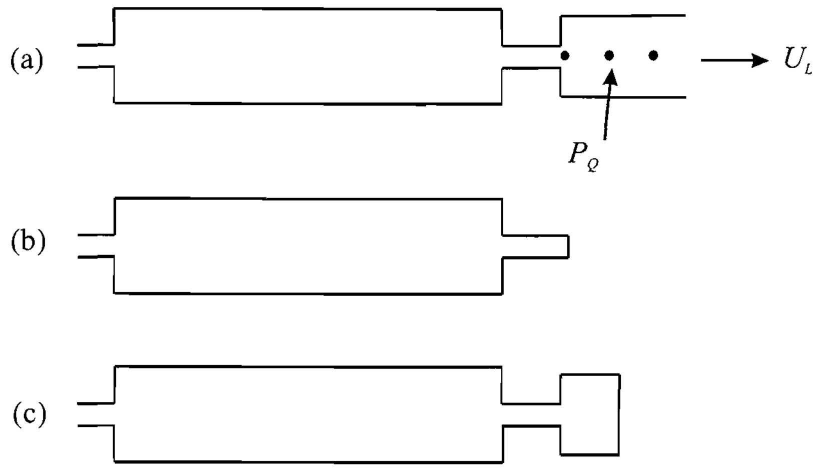

FIGURE 5.1: Model to illustrate calculation of H_{QL}^{P} when a noise source is located in front of a constriction in the vocal tract. U_o and P_s are the volume velocity and pressure source respectively. See text for explanation. Adapted from Stevens (1998).

1. *The acoustic medium is frictionless, homogenous, and at rest on the average.*

   ACTRA treats the air in the vocal tract as a saturated viscous heat-conducting fluid with flow having a time-averaged velocity u_o. The designation allows a closer simulation of the vocal tract environment, since the presence of water vapour introduces more variables into the calculation.

2. *The vocal tract walls are rigid.*

   ACTRA can model the walls as reacting surfaces which, in effect, attenuate or disperse the sound wave. More variables are introduced into the equation: wall mass per area, m, wall resistance per area, R and natural wall frequency ω_0. The values are adapted from Ishizaka et al. (1975).

3. *The effects of abrupt changes in area are negligible regardless of the degree of abruptness because the areas are approximate values and not actual cross-sectional areas.*

   ACTRA computes the end corrections at junctions when it calculates the TF.

5.1.1 Computation of Transfer Functions

Let us assume that the tube configuration shown in Figure 5.1(a) is a simplified version of the vocal tract area function from the glottis to the lips for a fricative consonant with the three dark circles downstream of the second constriction representing possible location of (noise) sources. In ACTRA, each noise source in the figure is represented by a pressure source, so that a hard termination is placed in the specified location when calculating the TF of the tract. The closed-end effect is shown in Figure 5.1(b) and (c) for two of the possible source locations.

The VTTF for the fricative is in effect the TF between the pressure source at the constriction to the lips. To calculate it, we must divide the TF from the glottis to the lips by the TF from the glottis to the source where the boundary condition at the source has the closed-end effect. Because of the difference in the boundary conditions of the two TFs, the TF from the glottis to the lips is known as the volume-velocity transfer function (VVTF) while the TF from the glottis to the source is the pressure transfer function (PTF). If we represent the VVTF from the glottis, G, to the lips, L, with H_{GL}^{V} and the PTF from the glottis, G, to the source, Q, with H_{GQ}^{P}, then the computed TF for the fricative can be expressed by:

$$H_{QL}^{P}(f) = H_{QG}^{P}(f) H_{GL}^{V}(f), \tag{5.1}$$

where $H_{QG}^{P}(f)$ is the reciprocal of $H_{GQ}^{P}(f)$.

In general, ACTRA calculates the VTTFs from the distribution of the partial pressures $p^{+}(f)$ and $p^{-}(f)$ of the positive- and negative-travelling component plane waves respectively. The $^{+}$ sign denotes a wave travelling from the glottis towards the lips and a $^{-}$ sign denotes a wave travelling in the opposite direction. Calculation of the component amplitude waves commences from the lips and moves back towards the glottis. The incident wave at the lips is assigned the value unity and the remaining amplitudes are systematically computed relative to this value until the glottis is reached, taking into account any changes in the dimensions of the vocal tract.

For computation of the VVTF, the reflection coefficient, R, at the lips, L, is set to a piston-in-baffle radiation impedance which results in a phase shift of 180° between p_{L}^{+} and p_{L}^{-}. The volume-velocity at this point is given by $U_L = (p_{L}^{+} - p_{L}^{-}) S_L / \rho_0 c_0$, where S_L is the cross-sectional area at the lips and ρ_0 and c_0 are the density of the medium and the speed of sound, respectively. The TF is the measure of the volume-velocity at the lips, L, with respect to the volume-velocity at the glottis, G, which is given by:

$$\begin{aligned} H_{GL}^{V}(f) &= \frac{U_L(f)}{U_G(f)}, & (5.2) \\ &= \frac{S_L}{S_G} \frac{(p_{L}^{+} - p_{L}^{-})}{(p_{G}^{+} - p_{G}^{-})}. & (5.3) \end{aligned}$$

where f refers to frequency and S_L and S_G are the cross-sectional areas of the lips and glottis respectively.

For computation of the PTF, the user specifies first the location of the intermediate pressure source in the vocal tract. At the pressure source location, Q, the reflection coefficient is assigned the value of unity (infinite impedance), consequently setting the partial pressures equal to each other, i.e., $p_{Q}^{+} = p_{Q}^{-}$. The pressure at point Q is then given by $p_Q = p_{Q}^{+} + p_{Q}^{-}$. The TF computed here is in effect, the measure of the volume-velocity at the glottis with respect to the acoustic pressure at point Q, given by:

$$H^{P}_{GQ}(f) = \frac{p_Q(f)}{U_G(f)}, \tag{5.4}$$

$$= \frac{\rho_0 c_0}{S_G} \frac{\left(p^{+}_{Q} + p^{-}_{Q}\right)}{\left(p^{+}_{G} - p^{-}_{G}\right)}. \tag{5.5}$$

It is clear that the TF of the fricative will have poles affiliated with the resonances of the whole tract from the VVTF and zeros affiliated with the back cavity from the PTF. Depending on the location of the pressure source, the poles and zeros of the back cavity may coincide with each other and cancel each other out, whereby we are left with only the poles affiliated with the front cavity and zeros affiliated with the region between the pressure source to the constriction. Thus, for the source location depicted in Figure 5.1 (b), complete cancellation of the back cavity poles and zeros may occur, whereas if the pressure source was located further downstream the constriction, as in (c), the result would be incomplete cancellation of the pole and zero pairs.

5.1.2 Computation of Far-Field Spectra

The Fourier transform of the sound pressure P_d, at a distance of d from the lips is given by,

$$P_d(f) = S(f)H(f)R(f), \tag{5.6}$$

that is, the product of a source function $S(f)$, a transfer function $H(f)$ and a radiation characteristic $R(f)$ (Fant 1960).

Therefore, to obtain predicted spectra comparable to the measured spectra, the TF of the fricative must first be multiplied to the source and radiation characteristic term. The radiation characteristic equation, which measures the sound pressure from the lips, is given by:

$$|R(f)| = \frac{\rho f}{d}, \tag{5.7}$$

where ρ is the density of air and d is the distance from the lips to the microphone (Shadle 1985). Here, ρ and d are assigned the values 1.18 kg/m^3 and 10 cm, respectively, because according to the experimental setup of the speech recordings described in Section 3.2, the speech utterances were recorded with a microphone placed 10 cm away from the subject's lips.

Variable	units	Description
`nE`		Total number of elements representing the area function. Single variable.
`elType`		The type of elements that were used starting from the lips. This is an `nE`×1 vector.
`elArea`	m^2	The unit areas of each element. This is an `nE`×5 matrix.
`elLength`	m	The unit lengths of each element. This is an `nE`×4 matrix.
`elHRad`	m	The hydraulic radius of each unit area of each element. This is an `nE`×5 matrix.

TABLE 5.1: ACTRA data file format and description.

5.1.3 User Input Data

ACTRA requires two types of inputs from the user: a data file and a parameter list. The data file consists of the vocal tract measurements previously derived in Chapter 4. The parameter list consists of a set of variables which correspond to the constituent properties of the vocal tract environment. We begin with a description of the data file.

The vocal tract measurements which make up the data file are modelled by a concatenation of predefined elements called `orifice`, `outlet` and `pipe`. These elements are shown in Figure 5.2. Each section of the vocal tract is modelled by a particular element. The `orifice` element models the reduction in the cross-sectional area and includes a provision for area changes at the junctions between each of its component units. The `outlet` element models any increment in the cross-sectional area and includes a provision to include side branches at the junctions. The `pipe` element models units of constant cross-sectional area. As shown in the figure, the variables S_1 to S_5 represent the unit areas of each element, while the variables X_1 to X_5 represent the unit length of each element. Each unit area has a corresponding hydraulic radius, denoted by the variables R_1 to R_5, not shown in the figure.

The breaking up of the vocal tract area function to a representation of simple tube configurations can be done manually, but for complex configurations, such as the area function, the process is error-prone and time-consuming. Thus, a function was created in MATLAB called `D4ACTRA` which systematically selects the most suitable element for each region. The output of `D4ACTRA` is in the format shown in Table 5.1. This is the required format of the data file for ACTRA. The function was tested meticulously by comparing the elements that it designated to those that were designated by hand for both simple and complex configurations.

The algorithm of `D4ACTRA` is as follows. The function reads in the first two to three units of the area function at any one time and selects the most suitable element to represent the

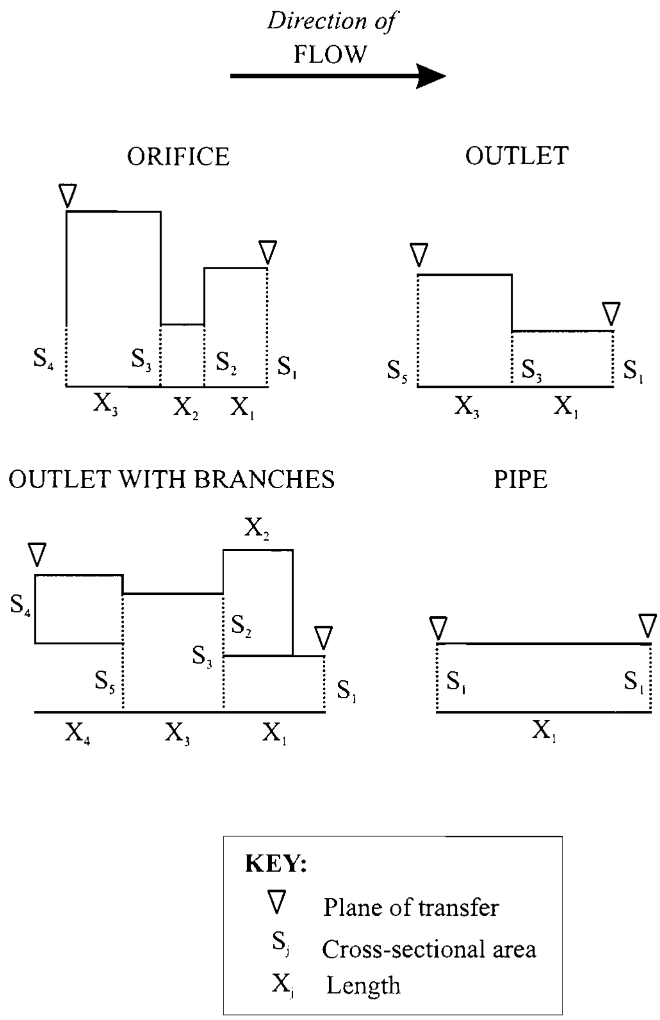

FIGURE 5.2: The predefined elements of ACTRA. Each section of the vocal tract is represented by one of the elements shown above. The variables S_1 to S_5 represent the unit areas of each element; the variables $X1$ to $X5$ represent the unit length of each element. Each unit area has a corresponding hydraulic radius (not shown) denoted by the variables R_1 to R_5. After Jackson (2000).

areas for that region. Once an element has been determined in elType, the variables S_1 to S_5 of elArea and X_1 to X_4 of elLength are assigned the corresponding values from the area function and length data, respectively. For the Blum transform, the interval between each successive grid plane is 5 mm, but for the Mermelstein-derived area, the intervals between successive grid planes at the vocal tract bend were not equal (see Section 4.4). The elHRad parameters were defined for each unit area of elArea. If, for example, the S_1 and S_2 variables of an element are assigned values, then the R_1 and R_2 variables of the element are also defined. Finally, the total number of elements representing the vocal tract configuration are counted and stored in nE.

The parameter list for ACTRA is shown in Table 5.2. It defines the constituent properties that allow ACTRA to simulate the vocal tract environment (or any other

Variable	Value	Units	Description
R	10000	kg s/m^2	Wall resistance per area
m	21	kg/m^2	Wall mass
ω_0	220	rad/s	Wall natural frequency
c_0	359	m/s	Speed of sound at 100% humidity
γ	1.3960		Ratio of specific heats at 100% humidity
p_0	760	Pa	Ambient Pressure
q_V	250	lit/s	Volume flow-rate
ρ_c	394	kg/m^{2}s	Characteristic impedance at 100% humidity
T	310	K	Humid and hot as in vocal tract
User preferences			
f_{max}	8000	Hz	Upper frequency
f_{min}	20	Hz	Lower frequency
nF	799		Number of frequencies $(1 + \frac{f_{\mathrm{max}}-f_{\mathrm{min}}}{\delta f})$

TABLE 5.2: The list of parameters specified in ACTRA.

environment) of the model and include other parameters such as the frequency range and frequency resolution of the output TFs. Specifically for the vocal tract environment, the various structures that form the vocal tract walls have different mechanical properties and are time-varying, but we assume that the defined values represent the whole of the tract and are time-invariant.

Volume flowrates of the vocal tract are known to differ for different speech sounds. The rate of airflow for fricative consonants produced at normal voice levels is usually between 200–500 lit/s (Stevens 1998). Higher flow-rates show higher amplitude levels in the spectrum while pole and zero frequencies remain somewhat unaffected (Shadle 1985). A flow-rate of 250 lit/s was employed for the acoustic predictions for all fricatives. Though there may be some differences in flowrates between the fricatives, the predicted spectra will be normalised to the same amplitude levels as the measured spectra prior to comparison, reducing the importance of the precise value of the flow-rate that was used.

Finally, the parameter list also requires that the user specify the frequency range and resolution of the TFs. Since fricatives exhibit more spectral energy in the higher frequencies, the frequency range is extended up to 8.0 kHz. For the planar-wave model to be valid, the diameter of the vocal tract must be less than a wavelength of the sound, so that the sound pressure is approximately the same at all points on the wavefront and the solution of the partial pressures depends entirely on the cross-sectional areas of the vocal tract (Stevens 1998). From the area measurements, the maximum diameter of the vocal tract for subject CS was estimated to be at 2.64 cm, which is substantially less than the 4.4 cm threshold for c_0 equal 353 m/s. Therefore, the planar-wave model is valid up to 8.0 kHz.

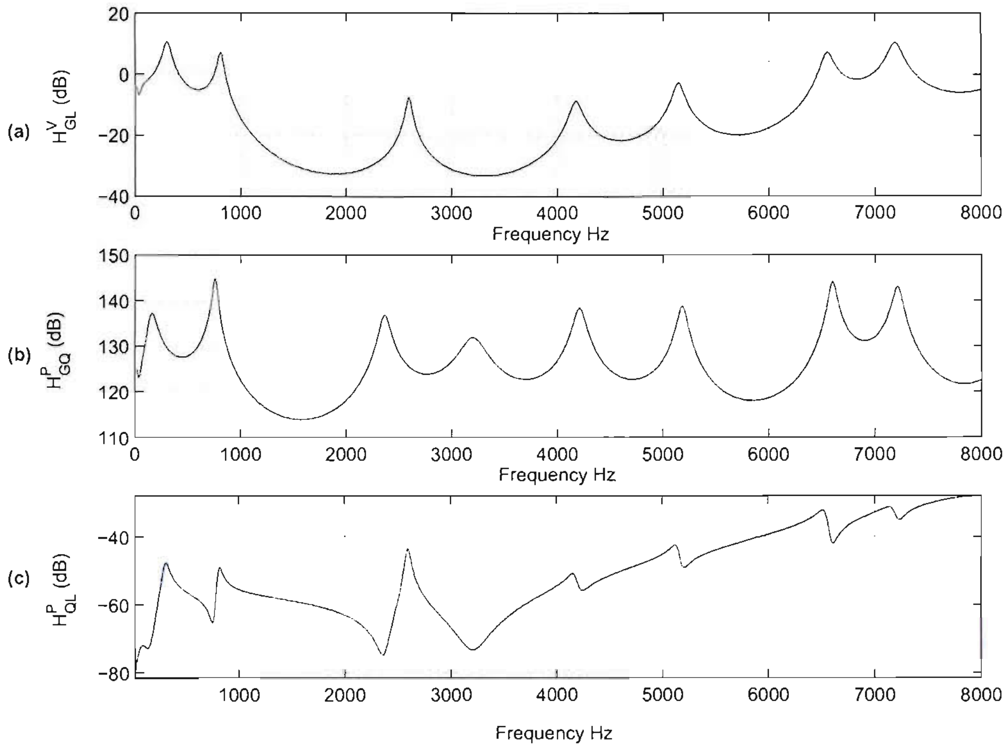

FIGURE 5.3: Example of the TFs generated by ACTRA for the area function of /s/. (a) H^V_{GL} is the VVTF of whole tract. (b) H^P_{GQ} is the PTF from source to glottis. (c) H^P_{QL} shown here is the product of H^V_{GL} with the reciprocal of H^P_{QL} and multiplied with the far-field response equation, $R(f)$, to obtain the predicted spectrum of the fricative.

5.1.4 Example: Deriving Transfer Functions of /s/

The following is an example of how the TFs for the area function of /s/ were calculated and how the final predicted spectrum is derived. Refer to Figure 5.3 for the output TF at each step.

1. The function D4ACTRA is used to produce the data file, M_1, for the whole of the area function (from lips to glottis).

2. Data file M_1 is processed by ACTRA to calculate the VVTF, H^V_{GL}, of the whole tract. The output TF is shown in Figure 5.3 (a).

3. The pressure source location is defined: the unit areas downstream of the determined pressure source location are removed from the area function.

4. The function D4ACTRA is used to produce the second data file, M_2, from the pressure source location to the glottis.

5. Data file M_2 is processed by ACTRA to calculate the PTF, H^P_{GQ}, of the whole tract. The output TF is shown in Figure 5.3 (b).

6. The PTF of the downstream H^P_{QL} is calculated by taking the product of H^V_{GL} with the reciprocal of H^P_{GQ}. By multiplication with the radiation characteristic term (Equation 5.7), the resulting spectrum is obtained shown in Figure 5.3 (c).

5.2 Summary of ACTRA

Several tests were carried out on ACTRA to ensure that it gives reliable predictions. The tests are presented in Appendix A and include simulation of various tube models and two slightly more complex configurations. The simpler tests include a uniform tube model with an intermediate pressure source, first with rigid, and then reflecting walls. Reflecting walls show perturbations in the lower frequencies and slight raising of the first formant, in agreement with data from the literature. The calculated error between the main resonances was found to be 4.6%. More complex tests include simulation of the mechanical flow duct model from Shadle (1985) and the area function from Narayanan and Alwan (2000) of an actual vocal tract of a subject sustaining /s/. The predicted spectra from both tests show resonance and trough frequencies that correspond well to the ones in the measured spectra. For the Shadle (1985) simulation, the error between the pole and zero frequencies was estimated to be 4.0%. For the area function of /s/, the predicted TF had similar characteristics to the spectrum presented in Narayanan and Alwan (2000). For example, the frequencies of the troughs rise as the pressure source location moves further up the tract.

ACTRA is able to model side cavities as separate ducts which are coupled to the main airway on one side and closed on the other using the `OUTLET` element. This enables modelling of the Blum-derived areas (quite) realistically and consequently, allows us to study the effects of including side branches on the fricative spectra and sounds. To verify ACTRA's ability in modelling side branches, the TF of a simple tube configuration with side branches was calculated using the `outlet` element (see Appendix A). The presence of a side branch introduces zeros into the system and the predicted zeros were compared to the calculated zeros. The calculated error was found to be less than 3.5%.

We conclude that ACTRA is well-suited for the purposes of calculating the TFs for fricative synthesis: it allows placement of an intermediate pressure source anywhere within the model and incorporation of the side branches into the model. The next chapter describes how superposition is used to derive the predicted spectra of a distributed source model. Subsequently, the differences between the measured and predicted spectra are compared qualitatively and quantitatively using the quantification procedure described previously in Section 3.7.1.

Chapter 6

Acoustic Predictions

Chapter 3 described the computation of the averaged PSD to obtain the measured spectra and subsequently, the analysis of the spectral characteristics of fricative steady-state. Chapter 4 described the derivation of the area functions of fricative articulation by adopting the techniques introduced by Mermelstein (1973) and Blum (1973). This chapter brings together the results of the previous analyses; first the area functions of Chapter 4 are used to derive the predicted spectra and then they are compared to the measured spectra of Chapter 3. ACTRA, the acoustic prediction software used to derive the spectra from the area measurements was introduced in the previous chapter.

The effectiveness of each area-derivation technique will be determined by:

1. comparing the measured and predicted spectra in terms of the amplitude levels and peak and trough frequencies and quantifying these differences using the degree of match measurement as described in Section 3.7.1; and

2. conducting a listening experiment, utilising synthetic fricative sounds derived from the predicted spectra, which will allow speculation on the perceptually important features of the spectra.

This chapter covers the first part. The analysis seeks to determine which area-derivation technique produces spectra that match best to the measured spectra and why, the effects of including side branches in the modelling (for Blum technique) and whether the predicted spectra match better to the measured spectra if side branches are included in the model. Acoustic pressure source locations are first considered and subsequently the predicted spectra for each fricative are derived.

6.1 Derivation of Predicted Spectra

Chapter 5 gave an example of how the TF of /s/ was calculated by ACTRA (Section 5.1.4). The final output of this procedure is the predicted spectrum (since it was multiplied with a radiation characteristic) of /s/ for a *single* point acoustic pressure source. In reality, the frication noise is not concentrated at a single point, but somewhat distributed downstream on the surface of the vocal tract walls and/or the teeth which act as an obstacle (Fant 1960; Stevens 1971; Narayanan and Alwan 2000). Therefore, the distributed source model was chosen to produce the synthetic fricatives. This required determination of the TF for a range of source locations in the vocal tract model; the result is several TFs for each fricative. The TFs are then superpositioned (to be described shortly) to derive the predicted spectrum for the given area function.

It must be emphasised that the type of noise at each point of the distributed pressure source is different: turbulent-flow induced sounds are generally represented by three types of noise sources known as the monopole, dipole and quadrupole (Lighthill 1954). However, implementation of different types of noise sources is outside the scope of this thesis, and although a distributed source was employed, the same white Gaussian noise source was applied at all pressure source points.

This section describes how the predicted spectra (since there are two) of each area function were derived and the process for determining the location of the pressure sources for each fricative configuration. The final part of this section describes how the predicted and measured spectra are normalised for comparison, so that the amplitude levels of the spectra are within the same range.

6.1.1 Distributed Source Model

The distributed source model for the fricative is a superposition of several predicted spectra, a predicted spectrum at a fixed distance from the lips having been given previously by Equation 5.6. If the source elements are independent of each other and the system being excited is linear, then the output of a multiple-source system, $P(f)$, can be expressed as:

$$P(f) = R(f) \sum_{i=1} H_i(f) S_i(f), \tag{6.1}$$

where $R(f)$ is the radiation characteristic term, $H_i(f)$ and $S_i(f)$ is the transfer function and source spectrum, respectively, for each pressure source location, i. Such an approximation may be used to represent the distributed nature of the source in terms of lumped non-interacting entities and therefore it is particularly useful for considering the effects of different source types. For random-natured processes such as fricatives, it is required to compute the modulus-squared of the predicted spectrum and the power spectrum of the noise source for superposition (Bendat and Piersol 1993).

As stated at the beginning of this section, the differences in the noise source representations are beyond the scope of this thesis and, for the synthesis to be described in the following chapter, the randn function in MATLAB was used to derive the source spectrum for all points of the pressure source. As the source spectrum, $S(f)$, is simply white noise, the shape of the predicted spectra is not affected by the source model in any way. Since this chapter focuses on comparisons between the predicted and measured spectra, we define here for predicted spectra the spectral envelope derived from the transfer functions and radiation characteristic term, so that, without the effects of noise, clearer comparisons can be made with the measured spectra. In other words, we factor out $S(f)$ from Equation 6.1 for now and concentrate only on the predicted spectra minus the source for the visual analysis. In Chapter 7, more explanation will be given of the source spectrum for derivation of the synthetic sounds.

6.1.2 Intermediate Pressure Source

The best combination of pressure sources for each area function was determined in an iterative process consisting of four steps: 1) predict; 2) select; 3) add; and 4) compare. These steps are described in more detail. First, the predicted spectra for approximately 10 pressure sources positioned 1.0 to 5.0 mm apart from each other along the anterior region of the vocal tract were derived. Following this, the spectrum for each pressure source location was plotted on a single graph and compared to the measured spectrum (of the given fricative). The peak and trough frequencies between the predicted spectra (for different pressure source locations) were carefully observed. Three to four spectra were selected, superpositioned and then compared again to the measured spectrum in a trial and error process, until the best combination of predicted spectra were found, with "best" defined as good matches in the amplitude levels (within certain frequency ranges) and the frequencies of the peaks and troughs.

For clarity, an example of the procedure described above is shown in Figure 6.1. Subplot (a) shows the three spectra of /s/ overlaid; the measured spectrum, the multiple-source predicted spectrum and the best matching single-source predicted spectrum. Subplot (b) shows the predicted spectra of various pressure source locations. In this particular example, the best multiple-source predicted spectrum was found to be a combination of pressure sources located at 0.2 and 0.3 mm downstream of the constriction and 4.0 mm upstream. The decision as to the best combination of spectra was based on the following observations of the features of the multiple-source spectrum: the significant peak at 2.8 kHz, the local minimum at 4.0 kHz and the match in the amplitude levels above 5.0 kHz. It is evident that the local minimum and maximum below 3.0 kHz do not present a good match to the measured spectrum. They are attributed to the back

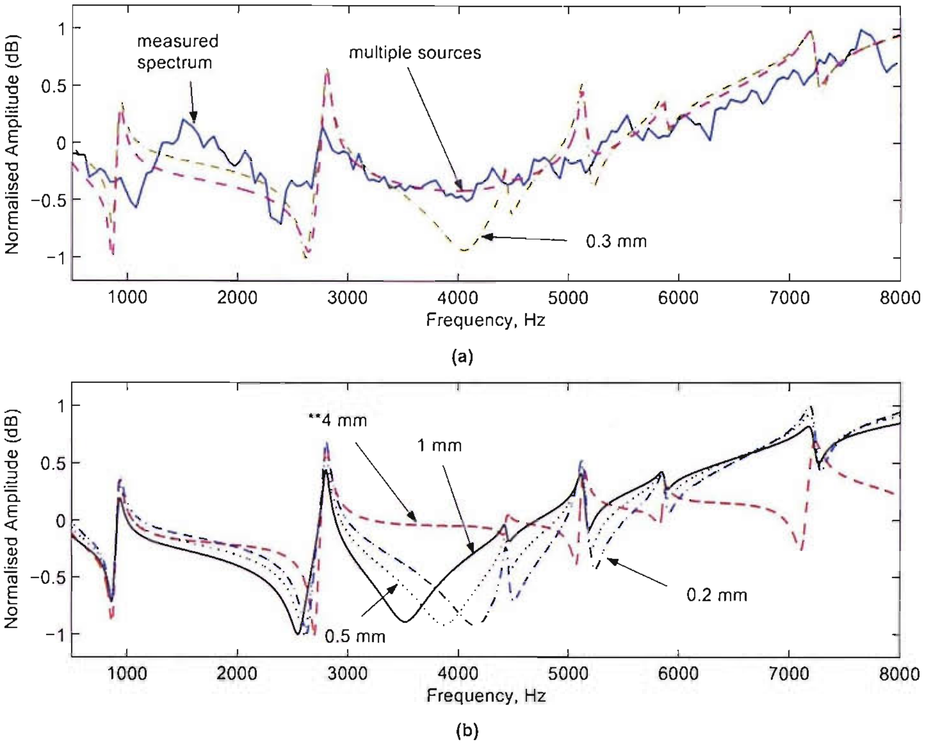

FIGURE 6.1: Determination of the pressure source combinations for /s/. (a) The measured spectrum for /s/ is overlaid with a single-source spectrum and a multiple-source spectrum that was determined to produce the best match to the measured spectrum, see text for explanation. (b) The various predicted spectra of /s/ of several pressure source locations, labelled with their corresponding distances from the end of the area of constriction. The ** denotes pressure source was located upstream from the end of the constriction region.

cavity (caused by incomplete cancellation of a pole and zero pair) and their frequencies are quite unaffected by the source location. Because of this, the differences in the lower frequencies are ignored and focus is primarily on the zeros, which are directly affected by the location of the intermediate sources. In light of this, it must be said that the system zeros, as well as the pole frequencies, are equally important in describing fricative spectra.

It should be noted that the distance from the constriction to the source is usually shorter than the front cavity length. Therefore, changing the location of the pressure source affects the frequencies of the zeros of the constriction-to-source location more than the front cavity poles (Shadle 1990). In essence, the zeros in a fricative spectrum can be divided into three groups: 1) a zero which always occurs at a very low frequency, resulting in low amplitude levels at the low frequencies for unvoiced fricatives; 2) a set of zeros which occurs at frequencies which are approximately equal to the resonances of the back cavity poles, thus effectively cancelling these poles; and 3) a set of zeros which occurs at frequencies related to the distance between the constriction and the source

Fricative	Location of pressure sources (mm)	
	from end of constriction	from lips
	Mermelstein technique	
/(a)f/	0.1, 1.0 upstream	0.1, 1.0
/(i)f/	1.0, 6.0 upstream	1.0, 6.0
/(u)f/	0.1, 4.0, 9.9 upstream	0.1, 4.0, 9.9
/θ/	3.0 downstream, 4.0 upstream	2.0, 9.0
/s/	0.5, 1.0 downstream, 4.0 upstream	4.0, 4.5, 9.0
/ʃ/	4.0, 16.0 downstream, 9.0 upstream	4.0, 16, 29
	Blum transform	
/(a)f/	1.5, 4.0, 6.5 upstream	1.5, 4.0, 6.5
/(i)f/	0.5, 4.0 upstream	0.5, 4.0
/(u)f/	0.5, 4.0 upstream	0.5, 4.0
/θ/	0.1, 1.0 downstream, 1.5 upstream	4.0, 4.9, 6.5
/s/	0.1 downstream, 2.0 upstream	4.9, 7.0
/ʃ/	1.0, 9.0 downstream	16, 24

TABLE 6.1: The locations of the pressure sources producing predicted spectra which best-matched the measured spectra. The distances are given with regards to the end of the constriction region (not the smallest constriction area) and from the lips. For nonsibilants, source locations are located upstream because the constriction occurs in the lip region.

location. The most obvious spectral peaks and troughs are thus the low frequency zero, the front cavity poles and the constriction-to-source zeros (Shadle and Scully 1995).

The locations of the pressure sources found to give the best predicted spectra for each fricative are shown in Table 6.1. The positions are described in terms of the distance from the end of the constriction, and, from the lips, so that the reader is able to relate them to realistic conditions where the frication source is located downstream of the constriction.

The pressure sources for nonsibilant fricatives were hardest to determine, because the constriction is located at the very end of the area function. Thus, possible locations for the pressure sources are limited. This is why the pressure sources are described as being located "upstream" of the constriction. They are, in fact, mostly located on the surface of the lips, which is similar to realistic conditions.

It was also observed from the table that the best location of the pressure sources differs according to the area-derivation techniques for the same fricative. This in itself is a significant observation; the area-derivation technique affects the possible locations of the pressure sources that are used to derive the predicted spectra. It does seem as if the Blum-derived spectra have the most realistic pressure source locations, since they are mostly located downstream or just slightly upstream as opposed to the pressure source locations of the Mermelstein-derived spectra.

There is the question of whether the pressure source locations may possibly compensate for inaccurate estimates of vocal tract geometry. For instance, in the case of vowels in LPC models, changing the shape of the source spectrum can produce relatively good versions of the vocalic sound even though the filter representation is inaccurate or completely wrong. For the fricative models here, it is speculated that the possibility of the source locations compensating inaccuracies in the filter representation is relatively small, first, because the source spectrum chosen was white noise which does not affect the shape of the final output spectrum and second, because the use of multiple sources (i.e., distributed source) will only "smooth" out zeros in the spectrum that are related to the distance between the constriction and the source location itself. Referring back to Figure 6.1, we can see that only the zero in the vicinity of 3.5 kHz is affected by the changes in the source locations whereas the poles and zeros at all other locations remain the same or only slightly affected. A clear example is the dislocated zero at 800 Hz which ideally should be located at 1.1 kHz: regardless of any combination of source locations, the zero stubbornly remains at the incorrect location.

Finally, it must be emphasised that the locations of the sources for fricative production are not deduced using ACTRA. On the contrary, it is an estimate of the best pressure source locations which will give the best prediction for the synthetic models. The choices are affected by several other uncertainties: the exact location of the teeth-airway boundaries, the forcing of 3-D shape to 1-D area function, the limitations of the modelling software because of the assumptions that it is based on and most importantly, the different area-derivation techniques.

6.1.3 Spectral Normalisation

Before comparing the measured and predicted spectra, normalisation was performed on the spectra, so that in the frequency range of interest, the amplitude of both types of spectrum would fall within the same levels. The normalisation process can be expressed by:

$$\hat{Y}(f) = \frac{Y(f) - \bar{y}}{\max\left(|Y(f)| - \bar{y}\right)}, \tag{6.2}$$

where $\hat{Y}(f)$ is the normalised predicted/measured spectrum and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} Y(i)$ is the average amplitude of the spectrum between the frequencies of 500 to 8 kHz.

Normalisation was applied to the spectra after taking the logarithm to base 10 of the spectrum. The reason for its application is mainly for the purposes of comparing the spectra, as no calibrated data were available for the recorded speech tokens. The procedure will cause the minimum or maximum amplitude of the spectra to fall between -1 and 1 dB. Thus, the amplitude values are arbitrary, but nevertheless allow for better comparisons qualitatively and quantitatively, because the amplitude levels are within the same range.

6.2 Spectral Analysis: Mermelstein versus Blum

The predicted spectra based on the Mermelstein and Blum area-derivation techniques are presented in this section. For each fricative, the predicted spectra of the two area derivation techniques will be compared to each other, and to the measured spectra. The area functions derived from the two area-derivation techniques and the differences between them have already been discussed in Section 4.4. Many of the arguments made here will be referenced to the area functions. To refer back to the area functions, turn to pages 72–75.

The results and discussion in this section will be divided into sibilant and nonsibilant fricatives. As mentioned in the previous section, the best source locations for nonsibilant fricatives were relatively harder to determine than for the sibilant fricatives. Recall from Section 3.5.1 that there were no distinct features that could be related to in the nonsibilant spectra. Pole and zero frequencies were observed to have a large variance in the steady state segments and the frequencies of troughs and peaks that fall within 200 Hz of the averaged measured frequencies are still considered a "good match".

Another important fact that the reader must remember is that the Mermelstein technique does not incorporate the areas of side branches unless the area is directly connected to the main tract within the measured grid plane (refer back to Figure 4.2, page 60). For consistency, the comparisons that were made between the Mermelstein- and Blum-derived spectra also excluded any side branches, which means that the side branches of Blum were excluded unless they were also connected to the main tract on the measured grid plane. The acoustic predictions that were derived with inclusion of the areas of side branches and comparisons with the measured spectra will be discussed separately in Section 6.3.

6.2.1 Sibilant Fricatives

Figure 6.2 (a) shows the overlaid measured (solid) and predicted (dashed) spectra of /s/, the predicted spectrum being derived using the Mermelstein technique; Figure 6.2 (b) shows the overlaid measured (solid) and predicted (dashed) spectra of /s/, the predicted spectrum being derived using the Blum technique. The pressure source locations for the Mermelstein-derived spectrum were 4.0 mm, 4.5 mm and 9.0 mm from the lips, and for the Blum-derived spectrum, they were located at 4.9 mm and 7.0 mm from the lips.

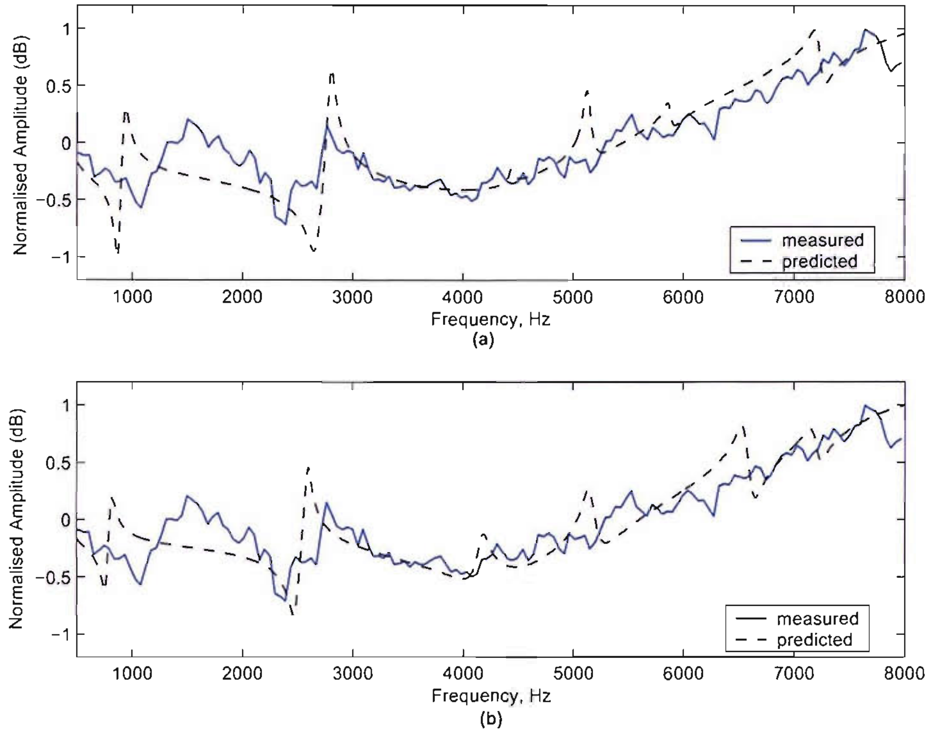

FIGURE 6.2: Measured and predicted spectra for /s/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived using the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived using the Blum technique without inclusion of side branches.

A comparison of the predicted spectra shows similarities in gross features; there is an incompletely cancelled pole-zero pair below 1.0 kHz, another between 2.0 to 3.0 kHz and a steady increase in amplitude levels above 4.0 kHz. The significant peak at 2.8 kHz of the Mermelstein-derived spectrum presents a slightly better match in frequency (than the significant peak from Blum-derived spectra) to the measured spectrum, but the amplitude levels of the Blum-derived troughs located at 800 Hz and 2.5 kHz are at similar levels to the measured spectrum. The first pole-zero pair of both the predicted spectra are clearly located at a very different frequency to the measured one. There are two likely reasons for this: either both techniques failed to represent the areas of the back cavity correctly, or subject CS was not in the same articulatory position during the speech and imaging sessions. The similarity of the spectra (and in the area functions shown on page 72) seem to support the latter suggestion.

Above 4.0 kHz, there are distinct differences between the two spectra: for the Mermelstein-derived spectrum, there are local maxima at 5.0 kHz, 5.8 kHz and 7.0 kHz; for the Blum-derived spectrum, there are local maxima at 4.2 kHz, 5.3 kHz, 6.5 kHz and 7.3 kHz. The major differences in the area functions between the two techniques are in the length of the constriction, which is longer by 5.0 mm, and the length of the back cavity, which is longer by approximately 15 mm for the Blum-derived area. The longer

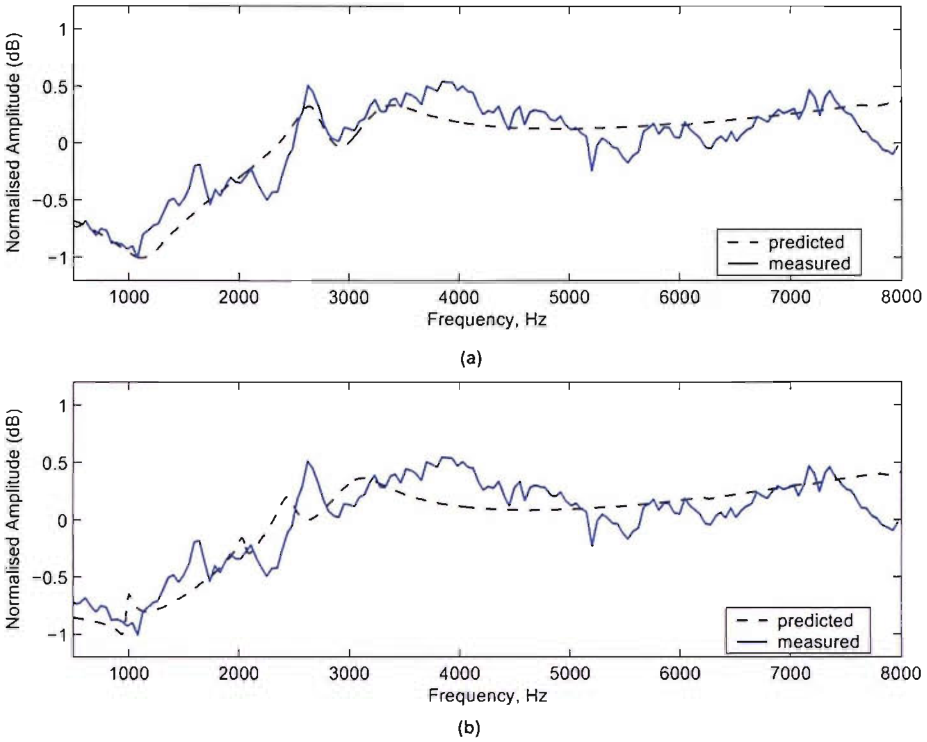

FIGURE 6.3: Measured and predicted spectra for /ʃ/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived using the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived using the Blum technique without inclusion of side branches.

back cavity length in the Blum-derived area may have slightly lowered the location of the pole-zero pair, while the longer constriction length and additional area measurement may have contributed to the differences in the formant locations above 4.0 kHz.

Figure 6.3 (a) shows the overlaid measured (solid) and predicted (dashed) spectra of /ʃ/, the predicted spectrum being derived using the Mermelstein technique and Figure 6.3 (b) shows the overlaid measured (solid) and predicted (dashed) spectra of /ʃ/, the predicted spectrum being derived using the Blum transform. The Mermelstein-derived spectrum was derived with pressure sources located at 4.0 mm, 16 mm, and 29 mm from the lips, and for the Blum-derived spectrum, they were located at 16 mm and 24 mm from the lips.

The distinct peak at 2.7 kHz is well matched in the Mermelstein-derived spectrum as well as the local minimum at 1.0 kHz. The Blum-derived spectrum also has a peak in the vicinity of 2.5 kHz. It is lower than the frequency of the measured peak but still within the acceptable range. The lowering in the frequency of the significant peak in the Blum-derived spectrum may have been caused by the longer length from lips to constriction in the Blum-derived area function.

The similarities observed in the gross features (e.g., peak in the vicinity of 2.4 kHz, higher amplitude levels above 3 kHz) of the predicted spectra are quite surprising given the significant differences between the area functions (refer to page 72). Recall that the Mermelstein-derived area showed very low area measurements for the laryngeal region whereas Blum showed very large measurements. This suggests that the cavities may have been decoupled during production of /ʃ/. This is likely, since /ʃ/ has the smallest constriction area (amongst all the fricatives). In other words, accurate measurements of the back cavity may not be important, because the poles affiliated with the cavity are almost, if not completely, cancelled by the system zeros, which is why the significant difference in the area measurements of the back cavity only creates small differences in frequencies below 3 kHz in the predicted spectra. Additionally, if decoupling did occur, then the pyriform sinuses may not have a significant role in the modelling of this fricative while the sublingual cavity may still be quite important.

In general, for the sibilant /ʃ/, neither predicted spectra exhibited the local minima at 2.3 kHz, 5.5 kHz, 6.5 kHz and 7.9 kHz found in the measured spectrum. It is unclear why this happened, but perhaps the thickness of the slices resulted in considerable loss of information. If one was to choose the best predicted spectra, the Mermelstein-derived version seems to have produced a better match to the measured spectrum, because of the match in the trough at 1.0 kHz, the peak at 2.5 kHz and the broad peak at 3.4 kHz—all of which were absent in the Blum-derived spectrum. However, the Blum-derived spectrum does show more detail below 3.0 kHz which may provide important cues to the listener as to the fricative's identity. This point is further considered in the next chapter.

6.2.2 Nonsibilant Fricatives

Figure 6.4 (a) shows the overlaid measured (solid) and predicted (dashed) spectra of /θ/, the predicted spectrum being derived using the Mermelstein technique and Figure 6.4 (b) shows the overlaid measured (solid) and predicted (dashed) spectra of /θ/, the predicted spectrum being derived using the Blum transform. The Mermelstein-derived spectrum was derived with pressure sources located at 2 mm and 9 mm from the lips and for the Blum-derived spectrum, they were located at 4 mm, 4.9 mm and 6.5 mm from the lips.

It is clear that the Blum-derived spectrum matches better to the measured spectrum than the Mermelstein-derived version. With the exception of the first pole and zero pair, displaced by approximately 200 Hz relative to the ones in the measured spectrum—but still within the accepted error bound—the pole and zero frequencies as well as the amplitude levels match up well to approximately 6 kHz. There is a mismatch between 6 and 7 kHz, but after 7 kHz the Blum-derived spectrum also produces the better match, as the amplitude level tapers down.

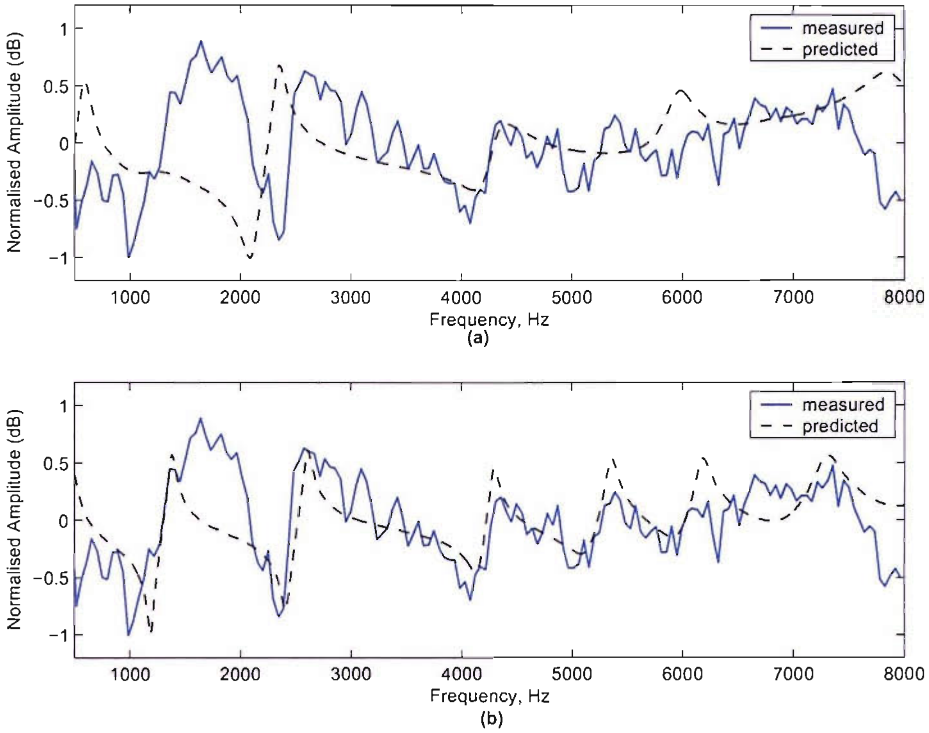

FIGURE 6.4: Measured and predicted spectra for /θ/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Blum technique without inclusion of side branches.

Referring back to the area functions of /θ/ (shown in Figure 4.13, page 73), the differences between the predicted spectra are likely to have been caused by the measurements of the laryngeal region. The Mermelstein-derived technique has produced very narrow vocal tract measurements, the major cause of difference between the area functions. Judging from the match between the measured and predicted spectra, the Blum-derived area seems to be a more accurate representation of the area function.

Figures 6.5 (a) and (b) show the overlaid measured (solid) and predicted (dashed) spectra of /(a)f/, the predicted spectra being derived from the Mermelstein and the Blum area-derivation technique, respectively. No combination of pressure sources predicted a satisfactory spectrum of /(a)f/ for the Mermelstein-derived area function mainly because of the limited area for possible source locations. The combination of sources to derive the predicted spectrum in (a) were located at 0.1 mm and 1.0 mm from the lips. For the Blum-derived spectrum, the pressure sources were located at 1.5, 4.0 and 6.5 mm from the lips.

Zeros in the lower frequencies are attributed to the resonances of the back cavity and their frequencies depend primarily on the location of the pressure source. If the pressure source is located slightly downstream, as in Figure 5.1(c), then the pole and zero frequencies

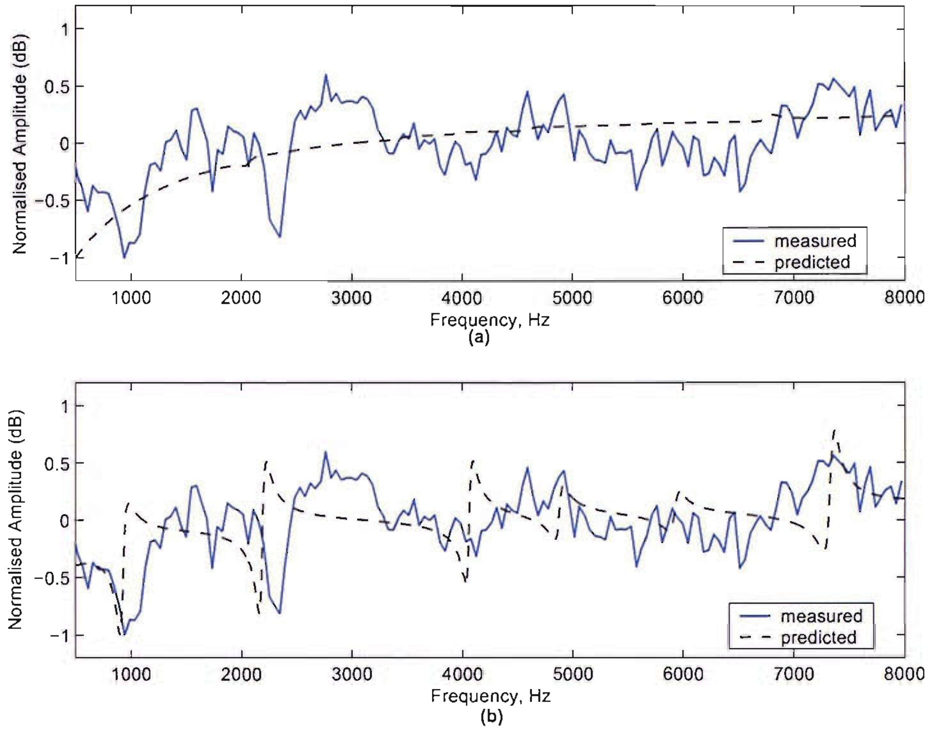

FIGURE 6.5: Measured and predicted spectra for /(a)f/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Blum technique without inclusion of side branches.

will be fairly distanced from each other, and the peaks and troughs of the system will be well-defined. But if the pressure source is located very near the constriction, such as in the configuration shown in Figure 5.1 (b), then the zero frequencies may well pair up with the frequencies of the front cavity and cause only small perturbations in the spectrum arising from incomplete cancellation of the pole-zero pairs (Stevens 1998).

For the Mermelstein-derived spectrum shown in Figure 6.5 (a), one probable explanation is that the poles from the glottis to the constriction coincide with the zeros from the glottis to the source, resulting in complete cancellation. This is because the pressure sources are located right at the constriction, similar to the example of Figure 5.1 (b). What might be expected to be seen in the spectra are the front cavity poles and the constriction to source zeros, but none were observed. In this aspect, ACTRA may have failed to model the fricative closely, primarily because the pressure sources were placed on the inner edges of the lips (0.1 and 1 mm), and, in this region, the theory of 1-D wave-propagation no longer applies. Several other pressure source locations were tried upstream of the constriction, but this did not improve the spectrum in any way. Downstream source placement was not possible being no longer in the model domain.

The Blum-derived area function, however, produced better predictions. Clearly, it is a better match to the measured spectrum than the Mermelstein-derived spectrum,

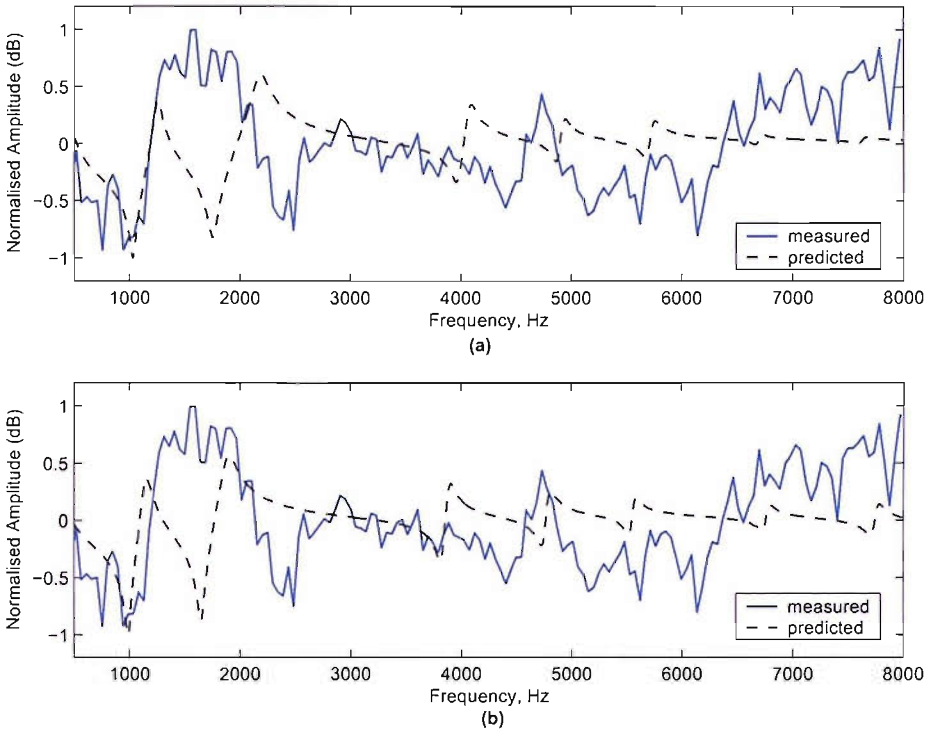

FIGURE 6.6: Measured and predicted spectra for /(i)f/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Blum technique without inclusion of side branches.

particularly at frequencies below 4.0 kHz. The first and second pole-zero pairs of the Blum-derived spectrum seems to be within 200 Hz of those in the measured spectrum, after which the amplitude decreases slightly, consistent with features of the measured spectrum. The explanation behind the main differences between the predicted spectra of /(a)f/ may be the shorter and smaller constriction of the Mermelstein-derived area function and the low area measurements in the laryngeal region, which may have ultimately reduced the effective length of the vocal tract and caused the complete cancellation of the back cavity poles referred to earlier.

Figures 6.6 (a) and (b) show the overlaid measured (solid) and predicted (dashed) spectra of /(i)f/, the predicted spectra being derived using the Mermelstein technique and the Blum transform, respectively. For the Mermelstein-derived spectrum, the pressure sources were located at 1.0 and 6.0 mm from the lips and for the Blum-derived spectrum, they were located at 0.5 and 4.0 mm from the lips.

The predicted spectra of both area-derivation techniques are almost identical, which is not surprising, as the area measurements (shown in Figure 4.15, page 74) show similar values. The major difference between the measured and predicted spectra is the lowered

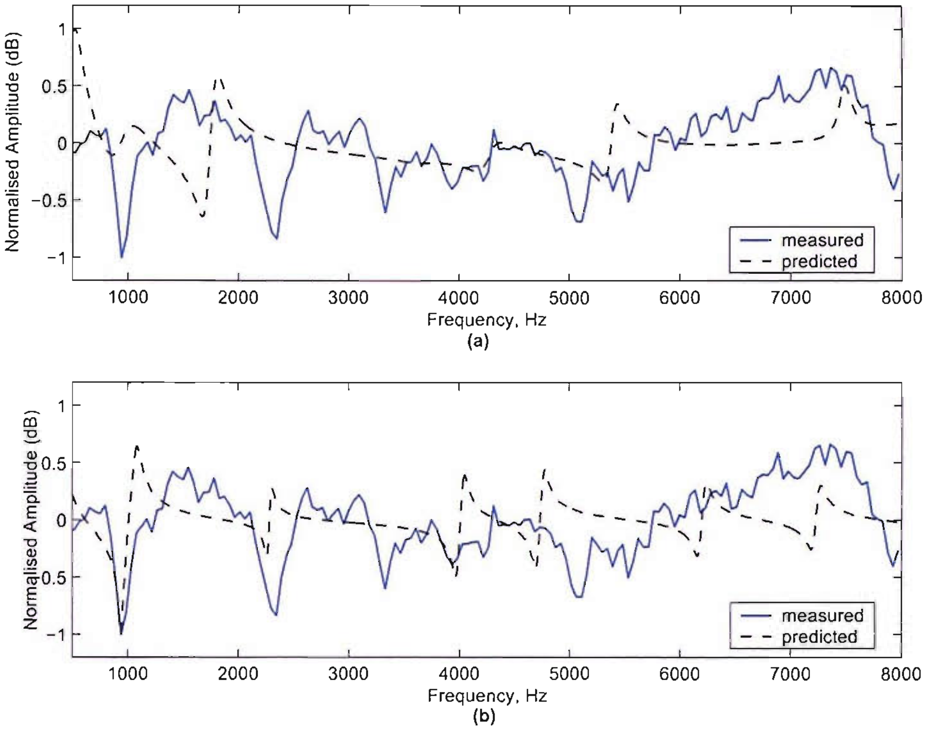

FIGURE 6.7: Measured and predicted spectra for /(u)f/. (a) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Mermelstein technique. (b) Measured spectrum (solid) and the predicted spectrum (dashed) derived from the Blum technique without inclusion of side branches.

frequency of the second trough, located at ±1.8 kHz in the predicted spectra and also of subsequent peaks and troughs up to about 6.0 kHz.

The mismatched second trough at ~1.8 kHz is slightly lower in frequency for Blum than for its Mermelstein counterpart. This is most likely caused by the additional length in the Blum area function, which exceeds the Mermelstein one by 20 mm. The similarity of the measurements of the area functions implies that the mismatch between the predicted and measured spectra at ~1.7 kHz was not caused by discrepancies in the area derivation but was most likely caused by the articulators being in different positions in the recording and imaging sessions.

Finally, Figures 6.7 (a) and (b) show the overlaid measured (solid) and predicted (dashed) spectra of /(u)f/ derived using the Mermelstein and the Blum area-derivation technique, respectively. The Mermelstein-derived spectrum was derived with the pressure sources located at 0.1 mm, 4.0 mm and 9.9 mm from the lips, while the Blum-derived spectrum combined sources located at 0.5 mm and 4.0 mm from the lips.

The Blum-derived spectrum appears to match better the measured spectrum based on the first two trough frequencies and the amplitude levels in this region. Apart from this, it seems as if neither area-derivation technique succeeded in producing a good match to the measured spectrum; the predicted spectra remain relatively flat above 3.0 kHz, characterised by low amplitude peaks which do not seem to match any particular peak in the measured spectrum. In general, the difficulty in making a good prediction is associated with the fact that the smallest constriction area lies in the lip region and that ACTRA may not give a close prediction, since the assumptions it is based on break down in this region.

Finally, it is evident that the predicted spectra of /(a)f, (i)f, (u)f/ are very different from each other. It has already been described in Section 4.4.3 how the effects of coarticulation were observed in the area functions of /f/ in different vowel contexts. It was also found from the quantification analysis of Section 3.7.2 how the measured fricative spectra of /f/ were most influenced by vowel context. Thus, it does not come as a surprise that the predicted spectra of /f/ in the three vowel contexts are very different from each other.

However, the different vowel contexts do not make the predicted spectra of the fricative match better the measured counterpart spectra. Vowel context /a/ shows the best prediction to the measured counterpart, but even then, only for the Blum-derived area function. It is possible that the effects of vowel context on the predicted spectra were enhanced due to inaccuracies in the area functions or as previously mentioned, simply a breakdown of the assumption of 1-D wave propagation, as the pressure sources were placed right at the end of the lips. In the next chapter, we speculate whether the differences brought on by vowel context are perceptually important to the listener.

6.3 Acoustic Predictions of Models with Side Branches

The Blum area-derivation technique derives the areas for the side cavities. This section investigates the acoustic effects from the inclusion of side cavities in the modelling by comparing the predicted spectra with and without side branches (derived using the Blum transform only) with the measured spectra. Four fricatives are analysed here:

1. the sibilant /s/, with inclusion of the pyriform sinuses, denoted as $/s/^{+}$;
2. the sibilant /ʃ/, with inclusion of only the sublingual cavity, and with inclusion of both the sublingual cavity and the pyriform sinuses, denoted as $/ʃ/^{*}$ and $/ʃ/^{**}$ respectively;
3. the nonsibilant /(a)f/ with inclusion of only the sublingual cavity and with the inclusion of both the sublingual cavity and the pyriform sinuses, denoted as $/(a)f/^{*}$ and $/(a)f/^{**}$ respectively; and
4. the nonsibilant /(i)f/ with inclusion of only the sublingual cavity and with the inclusion of both the sublingual cavity and the pyriform sinuses, denoted as $/(i)f/^{*}$ and $/(i)f/^{**}$ respectively.

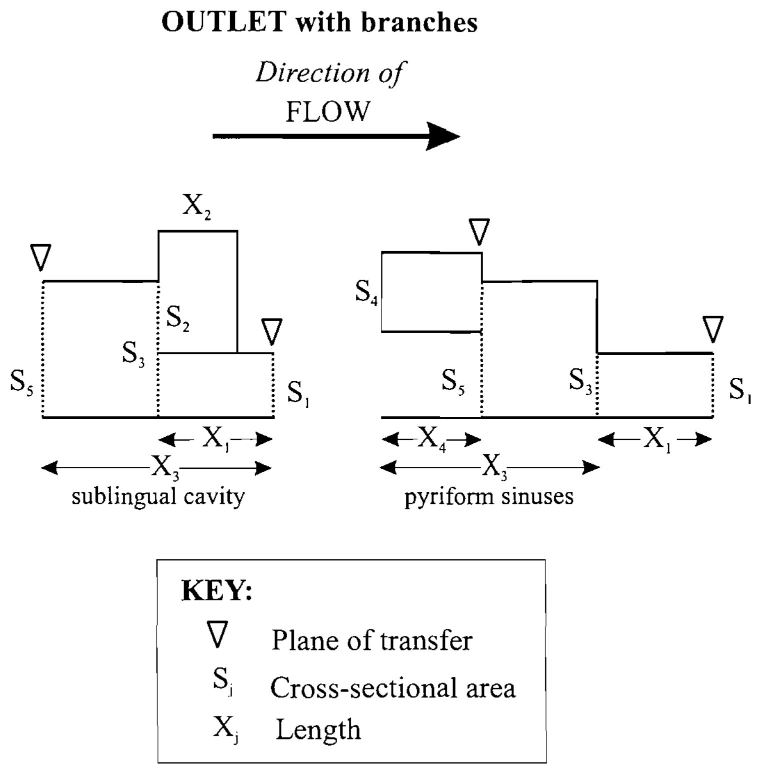

FIGURE 6.8: Modelling side branches in ACTRA. For modelling the sublingual cavity, only the right branch of the orifice element was utilised. For the pyriform sinuses, the left branch is utilised.

The locations of the intermediate pressure sources for each area function (with inclusion of side branches) were found using the procedure described in Section 6.1.2. Also, as previously stated in Chapter 4, while the tract bifurcates in the laryngeal region to form the two sections of the pyriform sinuses, for modelling, the areas of each individual sinus were added together and represented as one area. The hydraulic radius parameters were computed from the total area of the pyriform sinuses, $S_{\rm PS}$, using $\sqrt{\frac{S_{\rm PS}}{\pi}}$, which assumes that the cross-section is a perfect circle. The orifice element with side branches was utilised in different ways to model the sublingual cavity and the pyriform sinuses, as shown in Figure 6.8.

The predicted spectrum of /s/$^+$ overlaid with the measured spectrum is shown in Figure 6.9 (a). The spectrum was derived with only the inclusion of the pyriform sinuses as no sublingual cavity was evident during production. There are three major differences between the spectra of /s/$^+$ and that of /s/ without side branches (shown previously in subplot (b) of Figure 6.2): the decrease in amplitude of the local minima at approximately 4.0 kHz, a new peak introduced at 6.0 kHz and shifts in the peaks above 6.0 kHz. The differences in the upper frequency region (above 6.0 kHz) do not seem to create a better match to the measured spectrum, and the decrease in amplitude

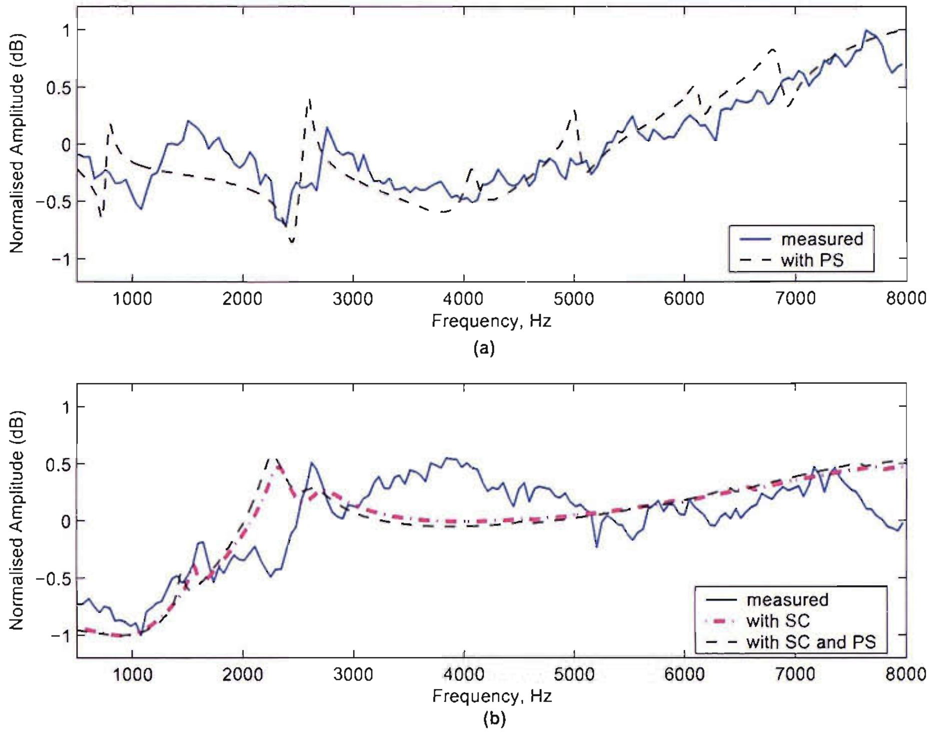

FIGURE 6.9: (a) Measured (solid) and predicted (dashed) spectra for /s/ modelled with inclusion of the pyriform sinuses (denoted PS). (b) Measured (solid) and predicted (dashed) spectra for /ʃ/ modelled with inclusion of the sublingual cavity (denoted SC) and with both the sublingual cavity and pyriform sinuses (denoted PS).

at around 4.0 kHz region seems to have introduced a larger difference from the measured spectrum compared to that for the spectrum derived without side branches.

The predicted spectra of /ʃ/* and /ʃ/** are shown overlaid with the measured spectrum in Figure 6.9 (b). In both cases, the amplitude level of the distinctive peak at ~2.3 kHz has increased and matches better the measured spectrum, no doubt because of the inclusion of the sublingual cavity (since it is apparent in both the spectra of /ʃ/* and /ʃ/**). The pole-zero perturbation at 1 kHz seen in the derived spectrum of /ʃ/ without side branches (previously shown in subplot (b) of Figure 6.3) has also disappeared with the addition of the sublingual cavity. For the predicted spectrum of /ʃ/*, there is a small peak at 1.5 kHz which matches the small peak in the measured spectrum. This peak is lowered by approximately 100 Hz in the spectrum of /ʃ/**.

The predicted spectra of /(a)f/* and /(a)f/** are shown overlaid with the measured spectrum in Figure 6.10 (a). With the exception of a slight shift in the pole-zero frequencies below 1.0 kHz, and between 4.0 to 6.0 kHz, the spectra of /(a)f/* and /(a)f/** seem to be almost identical. This implies that the pyriform sinuses may have only a very small role in the production of this fricative. When compared to the predicted spectrum of /(a)f/ without side branches, there seems to be some shifts in the pole and

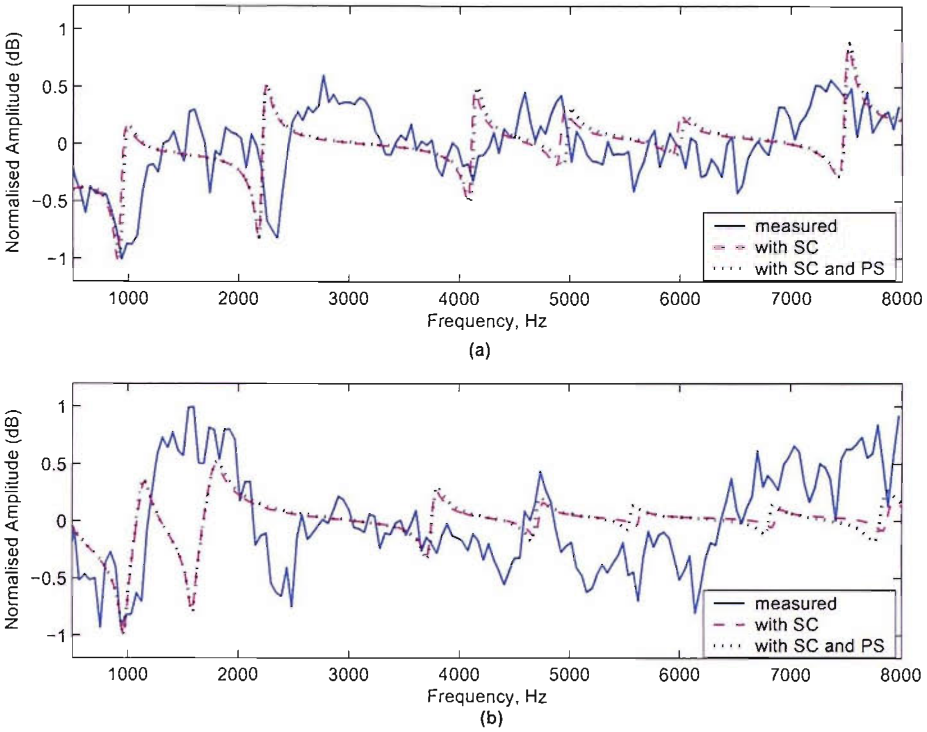

FIGURE 6.10: (a) Measured (solid) and predicted (dashed) spectra for /(a)f/ modelled with inclusion of side cavities. (b) Measured (solid) and predicted (dashed) spectra for /(i)f/ modelled with inclusion of side cavities. SC and PS denote the sublingual cavity and pyriform sinuses respectively.

zero frequencies by approximately 50 Hz, brought on by addition of the sublingual cavity and slightly more so with inclusion of the pyriform sinuses. The differences within the predicted spectra (including the one without side branches) are less than 50 Hz. If we consider the variation in the individual PSD spectra which was ±200 Hz, then the 50 Hz here can perhaps be negligible. However, we will still have to establish if these differences are perceptually important to the listener.

The predicted spectra for /(i)f/* and /(i)f/** are shown in Figure 6.10 (b) overlaid with the measured spectrum. As was the case for /(a)f/, the predicted spectra of /(i)f/* and /(i)f/** are almost identical, except for some slight shift in frequencies above 4.5 kHz. Again, we must establish if the small changes in the higher frequencies are perceptually important to the listener.

From the observation made here, addition of the side branches does little to improve the match between the predicted and measured spectra. Mainly, their inclusion seems to have lowered some peaks in the spectra, as observed in the case of /ʃ/, /(a)f/ and /(i)f/. If the cavities have little coupling as a result of the size of the constriction (as suspected for /ʃ/), then the pyriform sinuses will be of little importance because the back cavity poles will be completely cancelled. The sublingual cavity, however, is expected to have a higher degree of importance because it is located in the region where the frication noise is generated. Therefore, although addition of the side branches was observed to have little effect on the acoustic spectra, they may still be perceptually important to the listener.

6.4 Quantifying Differences in Spectra

The quantification procedure using dynamic time warping (DTW) and mean square error (MSE) measurements have already been described in Section 3.7.1 (page 48). Implementation of this technique to quantify the effects of vocalic context on the measured spectra of fricatives and calculation of the degree of match, D_m, (see Equation 3.6, page 50) has yielded results which are in agreement with findings from the literature. Consequently, we apply the quantification procedure to determine the D_m value between the measured and predicted spectra.

To summarise, the procedure consists of three steps:

1. the MSE between the predicted and measured spectra is measured to obtain E_B;
2. the DTW is applied on the two spectra and consequently, the predicted spectrum is warped;
3. the MSE between the measured and the warped predicted spectra is measured to obtain E_A; and
4. the D_m value is computed to quantify the degree of match between the spectra.

The higher the D_m value, the better the match between the spectra involved. For arguments supporting this statement, refer back to Section 3.7.1, where the motivation behind this technique is explained in more detail.

6.4.1 Spectral Envelope

Prior to the quantification procedure, a 20-point running average filter was applied on the measured spectra. The objective was to remove the effects of noise and obtain the spectral envelope. This will enable better comparisons between the measured and predicted spectra. The end effects of the filtering were not important: the full spectrum ranged from 0 to 24 kHz while only a segment from 500 Hz to 8.0 kHz was used for the analysis here. The choice of frequency range was explained earlier (Section 5.1.3). After smoothing, the measured spectra were normalised according to Equation 6.2.

Fricative	E_B	E_A	D_m
Mermelstein technique			
/s/	4.50	2.86	36.33
/ʃ/	2.05	1.17	43.02
/θ/	39.52	31.44	20.45
/(a)f/	8.92	8.31	6.85
/(i)f/	25.45	21.55	15.31
/(u)f/	20.69	17.58	15.03
		Average	22.83
Blum transform			
/s/	3.87	2.42	37.64
/ʃ/	3.86	2.70	30.00
/θ/	15.98	10.32	35.46
/(a)f/	11.94	6.16	48.42
/(i)f/	23.44	20.12	14.14
/(u)f/	21.18	16.30	23.05
		Average	31.45
Blum with branches			
/s/$^+$ (PS)	3.80	2.34	38.37
/ʃ/* (SC)	9.90	7.90	20.16
/ʃ/** (SC+PS)	8.30	6.36	23.66
/(a)f/* (SC)	11.85	6.81	42.52
/(a)f/** (SC+PS)	11.85	7.39	37.65
/(i)f/* (SC)	19.36	15.80	18.39
/(i)f/** (SC+PS)	19.84	15.97	19.51
		Average	28.60

TABLE 6.2: The degree of match, D_m, between the predicted and measured spectra. E_B and E_A denote the MSE measurements taken before and after DTW respectively. SC and PS denotes inclusion of the areas of the sublingual cavity and pyriform sinuses in the modelling respectively.

6.4.2 Quatitative Comparison of Spectra

Table 6.2 lists the measurements derived from the quantification procedure. The results are divided into three categories: 1) Mermelstein-derived; 2) Blum-derived; and 3) Blum-derived with side branches respectively. The second column lists the MSE measurements before application of the DTW, E_B, and the third column lists the MSE measurements after application of the DTW, E_A. The final column lists the degree of match, D_m, between the measured and predicted spectra.

The computed average of the D_m values of Table 6.2 for each group suggest that the Blum-derived area functions modelled without side branches produce predicted spectra that match better the measured counterpart than the Mermelstein technique. They also suggest that the role of side-cavities in fricative acoustics may not be very important, because the averaged D_m value for the models with side branches is not significantly different from those derived without. This will be further investigated shortly. In terms of the best- and worst-matched spectra, the Blum-derived /(a)f/ was observed to have the highest degree of match to the measured spectrum while the Mermelstein-derived /(a)f/ had the lowest degree of match. This observation is consistent with the features that were observed in the spectra, particularly in the case of the Mermelstein-derived /(a)f/ which was completely flat and very different from the measured spectrum.

In Figure 6.11, the D_m values were plotted against the cumulative distance of the DTW. Recall from Section 3.7.1 that the cumulative distance is directly proportional to E_A (given in column 3 of Table 6.2). As was discussed earlier (Section 3.7.1), the E_A value on its own is not a very informative parameter. However, for fricatives with the same D_m value, it is able to indicate the better-matched spectra. For example, consider the case of the Mermelstein-derived /θ/ and the Blum-derived /ʃ/*. The D_m value for each fricative is 20.45 and 20.16 respectively. Therefore, they both have the same degree of match to their measured counterparts. Now if we look at the E_A values, the Mermelstein-derived /θ/ has a measurement of 31.44 while the Blum-derived /ʃ/* has a measurement of 7.90, which is significantly lower. These values tell us that even though both spectra have the same degree of match to their measured counterparts after warping, the Blum-derived /ʃ/* is better than the Mermelstein-derived /θ/ because it required less warping to achieve the same degree of match.

Figure 6.11 plots the results presented in Table 6.2 in three different contexts. In subplot (a) they are plotted in terms of the area-derivation techniques, in (b) in terms of the fricative category, and in (c) in terms of inclusion or exclusion of the side branches. To interpret these figures, one must focus on the D_m values and then consider the cumulative distances. Thus, for the following discussion, let us assume that D_m values above 25 indicate a *good match* between the measured and predicted spectra and that values below this indicate a *poor match* between the measured and predicted spectra. In this context, the following interpretations are made for each subplot of Figure 6.11, with supporting arguments:

1. *The Blum-derived area functions produced better matches to the measured spectra than the Mermelstein-derived area functions.*
   53% of the Blum-derived spectra were good matches to the measured spectra, while only 33% of the Mermelstein-derived spectra produced a good match to the measured spectra.

2. *The predicted spectra of the sibilant fricatives matched better to the measured spectra than the predicted spectra of the nonsibilant fricatives.*
   71% of the good matches between the predicted and measured spectra were those of the sibilant fricatives /s/ and /ʃ/. For all degrees of match, good and poor, the sibilants show the lowest cumulative distance values (they are on the left-most

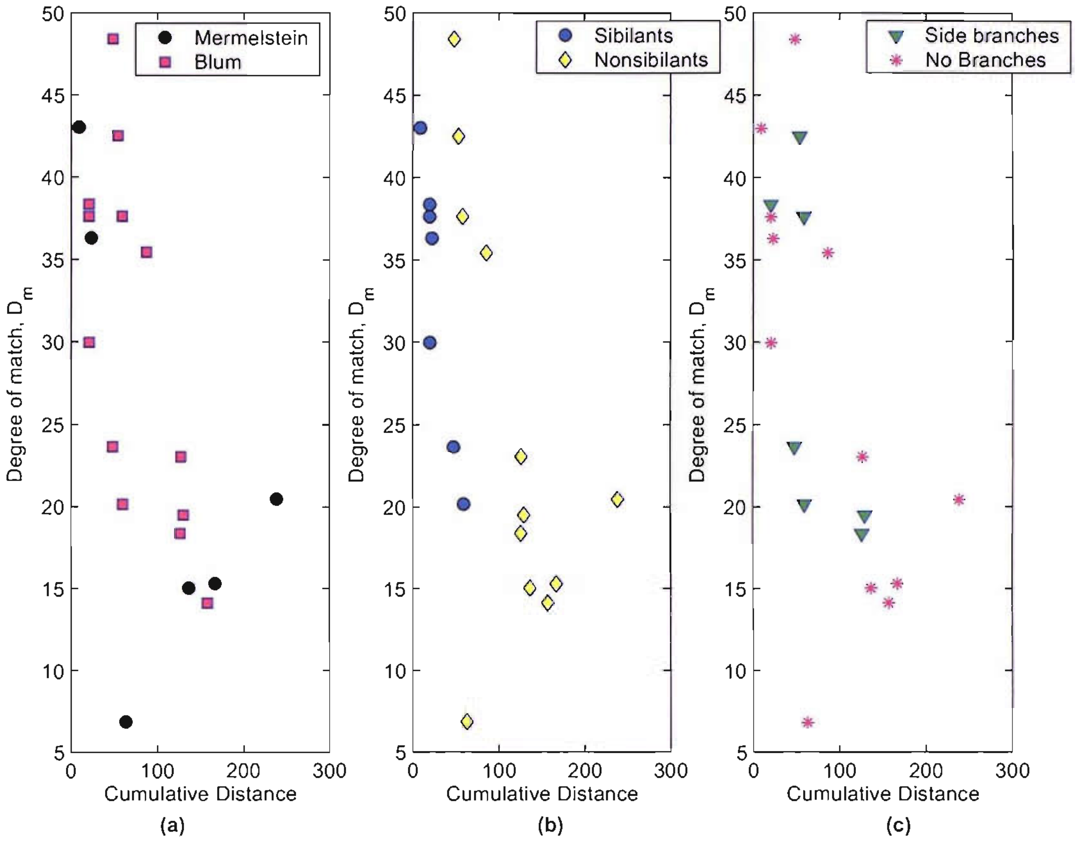

FIGURE 6.11: The results of the quantification procedure presented in terms of (a) area-derivation technique; (b) type of fricative and, (c) inclusion or exclusion of side branches. The cumulative distance is directly proportional to the E_A values shown in Table 6.2.

side of the figure), indicating that very little warping was applied to the predicted spectra to increase the match.

3. *The predicted spectra modelled with side branches did not increase the degree of match to the measured spectra overall.*

   42% of the predicted spectra derived with inclusion of the side branches succeeded in producing a good match to the measured spectra. However, 50% of the predicted spectra derived without side branches produced a good match to the measured spectra.

6.5 Discussion of Qualitative and Quantitative Analysis

The qualitative and the quantitative analyses have produced interesting results. The predicted spectra from the two area-derivation techniques show similarities in some cases (e.g., /s, ʃ,(i)f/) and differences in others (e.g., /(a)f, (u)f, θ/). The measured spectra of the sibilant fricatives were predicted well by both area-derivation techniques but apart from /(i)f/, the measured spectra of the nonsibilant fricatives were predicted better by the Blum-derivation technique, which suggests that the Blum technique is the best overall. With regard to the side branches, their inclusion had different effects on the acoustic features of each fricative. For /s/, addition of the pyriform sinuses did not significantly improve the degree of match, for /ʃ/ and /a(f)/ addition of the sublingual cavity and pyriform sinuses reduced the degree of match, while for /(i)f/, inclusion of the sublingual cavity and pyriform sinuses increased the degree of match between the measured and predicted spectra. These inconsistencies, along with the fact that the averaged D_m value with side branches in the model was lower than the averaged D_m value without side branches, suggest that side branches may not be very important to fricative modelling and the acoustic spectra can be more adequately modelled without them.

The predicted spectra are mainly dependent on the area functions, which in turn are dependent on the area derivation technique. If we recall the procedure for the Blum area-derivation technique, the major step was deriving the longitudinal axis of the vocal tract. This axis is perpendicular to each measured grid plane, so that each cross-sectional area represents the area that was encountered by the soundwave as it travelled along the tract. For the Mermelstein technique, the encountered area was calculated by multiplication with a cosine factor, which may not provide an accurate estimation of the area, particularly in the region of the side branches, where there are abrupt changes in the area. This is a potential reason why the Blum-derived area predicted better spectra than the Mermelstein-derived area.

Another important difference between the Mermelstein technique and the Blum area-derivation technique is the treatment of the junctions within the vocal tract. For the Mermelstein technique, the areas of the side cavities were only included if they were connected to the main airway on the measured grid plane. But for the Blum transform, the areas of the main tract adjacent to a junction were derived at an angle because of the presence of the side cavity. Therefore, although the areas of the side branches were also excluded in the first predicted spectra of Blum, the area measurements in the region of the side cavities were derived differently. The results of the quantification procedure have indicated that treatment of side branches—or more accurately, junctions—in this manner may be sufficient, as inclusion of the areas of the side branches as separate ducts coupled to the main airway did not improve the degree of match to the measured spectrum for all the fricatives. For /ʃ/ and /(a)f/ the degree of match decreased by approximately a third, which is relatively large, while for others, the degree of match was observed to increase. These results will be further compared to the results of the listening experiment which will be described in the next chapter.

Therefore, up to this point in the investigation, it is speculated that the Blum area-derivation technique produced a more accurate representation of the area function. This does not imply that the Mermelstein area-derivation technique is incapable of deriving good area functions. Indeed, the Mermelstein-derived spectra have the highest degree of matches for the sibilant fricatives. But the Blum transform was able to predict spectra which have a high degree of matches even for the nonsibilant spectra, particularly for /θ/ and /(a)f/. Previous studies noted the difficulty in deriving the acoustic spectra for nonsibilant fricatives, because of their strong sensitivity to lip shape (Shadle et al. 1995; Narayanan and Alwan 2000).

Before ending this discussion, another point must be observed about making comparisons: the match between the Blum-predicted spectra and the measured spectra was better in the lower frequency region than in the upper frequency region. In fact, for the sibilant fricatives, the match above ~3.5 kHz was mostly based on the amplitude levels, while peak and trough frequencies in this region either did not exist, as in /ʃ/, or were somewhat displaced, as in /s/. For the nonsibilants, the matches above 5.0 kHz were relatively poor, particularly for /(a)f/ and /(i)f/. The Blum-derived spectra were deemed better than the Mermelstein-derived spectra, mostly because of the better matches in the lower frequency region, which implies that the areas of the back cavity were derived more accurately by Blum. However, this also implies that either both techniques failed to derive the front cavity area accurately or, alternatively, both techniques were able to derive accurate representations of the front cavity area (since the area functions in the anterior region of the tract were fairly similar for both area derivation techniques), but the mismatch still occurred because:

1. The teeth areas were not extracted accurately from the vocal tract airway and consequently, contributed significantly to the mismatch in the higher frequency region of the predicted spectra, as was observed in most cases;

2. ACTRA has limitations as a result of various approximations and assumptions which the software uses to calculate the transfer functions, for example, when modelling sources on the lip region where the assumption of 1-D wave propagation may no longer apply; and

3. The far field transfer characteristic approximation, $R(f)$, which was multiplied with the predicted spectra may be unrealistic.

Finally, there is also the possibility that the subject was unable to assume identical articulatory positions for the MRI and audio recording sessions. This seems the most probable cause of mismatch between the measured and predicted spectra of /(i)f/ and /s/. The measurements of the area functions of both techniques were similar but the predicted spectra differed significantly from the measured spectra because of the lowered frequency of the second pole-zero pair. The findings here are not conclusive and the predicted spectra will be investigated further in the next chapter.

6.6 Summary and Conclusions

This chapter described how superposition was used to derive the predicted spectra of a distributed source model for each fricative. The predicted spectra, derived through the Mermelstein and Blum area-derivation techniques, were first compared with each other and then with the measured counterpart in a qualitative and quantitative analysis. The Blum-derived spectra were modelled first without side branches and later with side branches to study the effects of including the areas of side cavities on the acoustic spectra.

The quantitative analysis measured the degree of match between the measured and predicted spectra. The results of this analysis were in agreement with the findings that were made when the spectra were compared qualitatively, that is, the better the observed match between the predicted and measured spectra, the higher the D_m value. For example, the Mermelstein-derived spectrum of /(a)f/ was flat and did not exhibit any features that were common to the measured spectrum. The quantification procedure resulted in the lowest D_m value of 6.85% in this case.

Up to this point in the investigation, a decision was made that the Blum-derived area functions were able to predict the best-matched spectra overall while the Mermelstein technique produced the best-matched spectra for sibilant fricatives only. Two potential reasons were identified for the better matches of the Blum-derived spectra: the use of the longitudinal axis for the placement of the grid planes and the treatment of junctions within the tract.

The role of the sublingual cavity and the pyriform sinuses was unclear because for some fricatives, the degree of match improved with their inclusion (i.e., /s, (i)f/), while for others, it worsened (i.e., /ʃ/, (a)f/). The effects of including both the sublingual cavity and the pyriform sinuses into the model were similar to the effects obtained if only the sublingual cavity were included in the model. This implies that the pyriform sinuses are of lesser importance. In general, the addition of side branches resulted in small changes to the predicted spectra: an additional peak at 6 kHz for /s/, an increase in amplitude level of the peak at 2.3 kHz and shifts within 100 Hz of some peaks/troughs for /(a)f/ and /(i)f/. Although these changes are seemingly insignificant, the listening tests will determine whether they are perceptually important to the listener. It is possible that side branches may not have a significant role in fricative production, particularly if there is very little coupling activity and the back cavity poles are completely cancelled. But there is insufficient evidence here to conclude this firmly.

In the next chapter, the predicted spectra are multiplied with noise sources to derive the synthetic fricative sounds. Subsequently, a series of listening tests, conducted on 18 listeners, will be described. The results here along with the results of the listening experiment will allow us to study and speculate on the perceptually important features of the acoustic spectra of fricatives.

Chapter 7

From Images to Sounds

In the previous chapter, the predicted spectra for each fricative were derived from the area measurements and subsequently compared to each other and to the measured spectra in a qualitative and quantitative analysis. In general, the results suggest that the Blum transform predicted the best-matched spectra overall, while the Mermelstein technique predicted the best-matched spectra for only the sibilants. The importance of including side branches in the modelling is as yet unclear; in some cases their addition improved the degree of match, while in others, it did not. One possible suggestion is that that their role may not be very important for fricative production. This chapter hopes to clarify further the findings of the previous chapter. The predicted spectra will be used to derive synthetic fricative sounds and a series of listening tests utilising these sounds will be conducted on 18 listeners.

This chapter first describes how the synthetic fricative sounds were derived, then introduces the term "pitch quality", a perceptual phenomenon that gives the impression to the listener that each fricative sound segment has a different pitch or tone, although no voicing source is involved. Subsequently, the listening experiment is described, which is divided into five separate tests, each with a different objective, and which utilises different combinations of synthetic and natural speech segments. The next section describes how statistical tests were applied to the results of the listening experiment to determine their significance, before final analysis of the results and derivation of conclusions. In this chapter we sought to achieve two out of the three major objectives defined in Chapter 1, namely to consider the perceptual features of the acoustic spectra that are important to the listener and may have influenced decisions made during the tests, and to determine the importance of including the side branches in the model and the perceptual effects of their inclusion on the listener.

7.1 Derivation of Synthetic Sounds

The derived synthetic speech sound is a product of a transfer function, $H(f)$, multiplied by the radiation characteristic term, $R(f)$, with a source spectrum $S(f)$, as defined by Equation 5.6. The spectral envelope of the predicted spectra, computed by taking the product of $H(f)$ with $R(f)$, was already discussed in Chapter 6. Thus, this section will proceed directly to the derivation of the source spectrrum, $S(f)$.

It must be recalled that the maximum frequency of the predicted spectra was 8.0 kHz. The frequency resolution of the predicted spectra was 10 Hz (see ACTRA parameter list, Table 5.2, page 89). If a sampling frequency, F_s, of 16 kHz was used, the duration of the sound that can be derived from the predicted spectrum would be only 0.1 s. Thus, to obtain sounds of a sufficient duration, the overlap-add technique described in Rabiner and Schafer (1978) was adopted for the synthesis. The technique requires an overlapping of several waveforms that have been previously windowed. The relative displacement between two waveforms only occurs in the region where the signal is tapered from the effects of windowing. Thus, it is overlapped in such a way that, when the amplitudes of the waveforms are added, the maximum amplitude in the overlapping region is no more than the maximum amplitude of the original waveform. For clarity, an example is shown in Figure 7.1. Subplots (a) to (d) show four synthetic waveforms of the fricative /ʃ/. The beginning and end of each waveform are tapered from the effects of windowing. The four waveforms are then summed, with a relative displacement of 50% between segments, to form the final waveform shown in (e). Since each waveform has a duration of 0.1 s, the final sound segment of four waveforms will have a duration of 0.25 s.

Thus, to obtain the synthetic sound for each fricative with a duration of 0.25 s, four synthetic waveforms must be derived to apply the overlap-add technique. For the creation of each waveform, a new set of random numbers was generated. Thus, four different noise spectra were used. To prevent variation in the sound between the synthetic segments from the effects of source variation, the same four sets of random sequences were used to derive the synthetic sounds for all the fricatives.

The following steps were taken to generate a single synthetic waveform. For the representation of noise, a set of random numbers was first generated using the `randn` function in MATLAB. Following the usual notation, we refer to the noise signal as $s[n]$, with n denoting each sample point. The fast Fourier transform (FFT) of $s[n]$ was first computed to obtain the spectrum:

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-j(2\pi/N)kn} \quad k = 0, 1, \ldots, N-1. \tag{7.1}$$

where $N = 1600$ sample points. The sampling frequency, F_s of the recorded speech corpus was 48 kHz, but the acoustic predictions were only derived up to 8 kHz. This

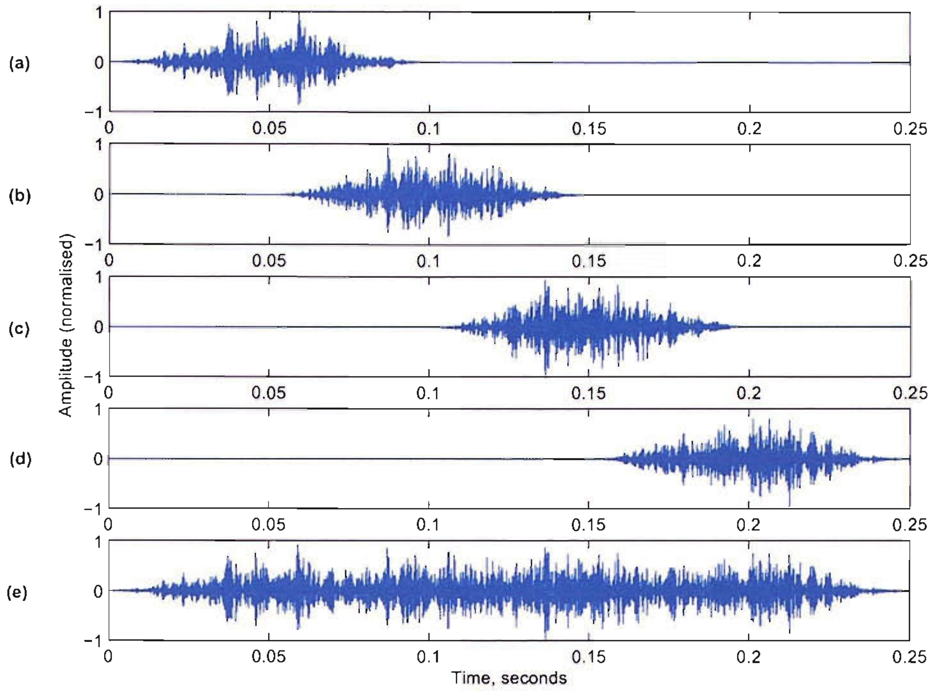

FIGURE 7.1: The overlap-add technique applied on the waveform of synthetic /ʃ/. Subplots (a) to (d) show the four waveforms derived using the predicted spectrum but different source models. Each waveform was Hanning windowed and added together with a relative displacement of 50% to derive the final waveform shown in (e). To avoid variation between the fricative sounds from the effects of noise, the source model was computed from the same set of random numbers across the fricatives.

limitation allows a maximum sampling frequency of 16 kHz (which also means that the recorded speech signals must be down-sampled in order to make a fair comparison between the synthetic and natural speech sounds). Thus, with F_s fixed at 16 kHz, taking the FFT with $N = 1600$ sample points will produce a source spectrum with a maximum frequency of 8 kHz and a resolution of 10 Hz, the same frequency range and resolution as the predicted spectra.

The power spectrum of the source was then computed according to:

$$\bar{S}[k] = \frac{1}{N}|S[k]|^2 \quad k = 0, 1, \ldots, N-1, \tag{7.2}$$

and subsequently multiplied by a Hanning window that tapers 20% of the sample points at the beginning and end of the spectrum, while the remaining 60% of the mid-section of the window is kept at a constant level of 1. This step is important because the effects of "ringing" (known as the Gibbs phenomenon, after Gibbs and Wilson 1960), which appears after transforming the noise spectrum back to time domain, is significantly reduced. Ringing appears as high amplitudes in the beginning of the waveform and is heard as clicks when the synthetic sound is played. The source spectrum is subsequently multiplied to the spectral envelope of the predicted spectra to obtain $P(f)$.

Finally, the inverse FFT of $P(f)$ was computed for $N = 1600$ sample points:

$$p[n] = \frac{1}{N}\sum_{k=0}^{N-1} P[k]e^{j(2\pi/N)kn} \quad n = 0, 1, \ldots, N-1, \tag{7.3}$$

where $p[n]$ is the synthetic waveform in the time domain. For $F_s = 16\,\text{kHz}$, the result is a sound segment of 0.1 s in length. Thereafter, the derived waveforms $p_i[n]$ for $i = 1, 2, 3, 4$ are processed according to the overlap-add technique, as described above. Thus, the final product is a synthetic sound with a duration of 0.25 s.

7.2 Processing of Natural Sounds

To enable a fair comparison between the synthetic and naturally-uttered fricative segments, the recorded segments must also have a sampling frequency of 16 kHz, be processed using the overlap-add technique and have a final duration of 0.25 s. To achieve this, the signal was first down-sampled from 48 kHz to 16 kHz. Subsequently, four consecutive segments were extracted from the steady state section of a single utterance, each segment consisting of $N = 1600$ sample points. From here, the four segments were windowed and summed together with a relative displacement of 50% as was previously described for the synthetic fricatives (again refer to Figure 7.1 for the example). The result is a sound segment with a duration of 0.25 s for each fricative.

Finally, the natural waveform was put through a 2nd order high-pass filter with a cut-on frequency of 40 Hz. The objective was to remove some ambient/microphone noise that occurred while the recordings were made (which can clearly be heard in the "silent segments" when the waveform is played). For consistency, the synthetic segments were also filtered in this manner.

7.3 Listening Subjects and Experimental Setup

There were 18 listeners who participated in the listening experiment: 16 were male and 2 were female. All the listeners were native English speakers. The age range of the subjects was 20 to 48 years old with a mean age of 29 years. The listeners were staff or students at the University of Southampton and originated from various parts of the UK. Before the experiment, the subjects were aware that they would be required to listen to some sounds but they did not know and were not informed that the sounds they would be hearing were actually (synthetic and natural) sounds of speech, particularly fricatives.

The experiment took place in an anechoic laboratory on the premises of the School of Electronics and Computer Science at the university. The subjects were required to wear

SONY headphones (model MDR-XD200) during the test so that the sounds could be heard as clearly as possible. The experiment consisted of five tests overall, but a short preliminary test was given to ensure that the listeners understood the instructions and knew what to expect. All the tests were given in a single session with each session lasting approximately 40 minutes per person. The listeners had no control over the sounds that were presented and were provided with only a pen and the necessary forms on which to mark their responses. Prior to the start of the fifth test the listeners were asked to indicate (i.e., by circling a YES or NO) if they had any training in phonetics and/or if they had some experience of listening to synthetic speech. None of the listeners had any such experience.

7.4 Pitch Quality of Fricative Sounds

The segments of sound that were derived from the predicted spectra and those from the processed natural recordings can be heard at the author's website `http://www.ecs.soton.ac.uk/~kss01r/sounds/`. A total of 25 sounds was used in the listening experiment:

- the natural segments of /(a)f, (i)f, (u)f, θ, s, ʃ/;
- the Mermelstein-derived segments of /(a)f, (i)f, (u)f, θ, s, ʃ/;
- the Blum-derived segments of /(a)f, (i)f, (u)f, θ, s, ʃ/ with no side branches; and
- the Blum-derived segments with inclusion of the pyriform sinuses for /s/$^+$, inclusion of sublingual cavity for /ʃ/*, /(a)f/* and /(i)f/* and inclusion of both sublingual cavity and pyriform sinuses for /ʃ/**, /(a)f/** and /(i)f/**.

No attempt will be made to describe the sounds here, as this will be discussed in depth according to the results of the listening experiment. However, before continuing further, it is a good idea to introduce the term "pitch quality", since it will be used frequently to refer to the pitch of the fricative sounds. It is emphasised here that in this context, the pitch quality of a sound does not carry the same definition as the perceived pitch in voiced sounds, which is quantified by the fundamental frequency. Instead, it is associated with the location of the spectral peaks in the noise that results in different "notes" between the sounds and it is found to be significantly perceived by the listener.

An example can be seen in the predicted and measured spectra of /ʃ/ (shown previously in Figure 6.3, page 100). There is a distinct peak at 2.7 kHz in the measured spectrum and the Mermelstein-derived spectrum. In the Blum-derived spectrum, the peak is lowered to 2.5 kHz. What the listeners perceive is that the pitch quality of the natural /ʃ/ is similar to that of the Mermelstein-derived /ʃ/ sounds. While the Blum-derived /ʃ/

still sounds like an /ʃ/ sound, it carries a different pitch quality than that of the natural sound.

What has brought this to our attention is the feedback from the listeners of the listening tests themselves. Having completed the listening experiment, some listeners confessed that, since they were unaware that the sounds they were hearing were actually fricative speech sounds, their decisions for the first four tests had been influenced by what they called the "tone" or "pitch" of the sound. In most cases, these listeners had strong musical backgrounds and were sensitive to the sound pitch.

7.5 Description of Listening Tests and Preliminary Results

The listening experiment consisted of five separate tests. The first four tests were "two-alternative forced-choice" (2AFC), while the fifth test used a rating scheme. The following subsections describe the tests in detail. Note that the results of each listening test will also be given in this section. These results are not final; they will be further analysed using statistical tools in the following section.

The tests were given with the order randomised across subjects. Once each test started it could not be paused. Therefore, there was no interaction between the listening subject and the experimenter during the course of the test. In fact, after the first preliminary test (which was given to ensure that the listener understood the instructions and knew what to expect of the experiment), there was no interaction even between tests; it was conducted in such a way that the listeners would read the instructions and gave an indication of when they were ready to start. Once a test was complete, they automatically moved on to the next test until all the tests were done.

It must be emphasised that in the first four tests, the listeners were not aware that they would be hearing synthetic and natural segments of speech. It has been known that outcomes of perceptual experiments differ when listeners were told they were actually listening to speech sounds rather than just *sounds* (i.e., Best et al. 1981; Johnson and Ralston 1994). However, the design of the experiment which involves short bursts of noise segments, template-matching, limited decision-making time and forced-choice conditions (which will be described in more detail in the oncoming sections), allows us to speculate that there would not be a major difference in the outcome of the experiment whether or not the listeners were informed that the sounds were actually speech. Unless, of course, more detailed information were given about the sounds that were presented, for instance *"sound one is fricative X, sound two is fricative Y and sound three is fricative X. Which category does sound X belong to?"* which would negate the task they were undertaking altogether.

It was noted from the experiment conducted by Harris (1958) (see subsection 1.4.3), that the listeners were not able to identify correctly the nonsibilant fricatives without the information contained within the vowel-to-fricative transitional period. The experimental setup in this case is different to that of Harris's because the listeners tasks are now template-matching rather than simple categorisation of the sound. Thus, the experiment was designed in such a way that it will test the listeners' ability to match two sounds that may or may not sound alike but are within the same category. Therefore, even though Harris (1958) observed that the nonsibilant sounds cannot be identified correctly by the listener when asked to choose between /f, θ, s, ʃ/, there is still the possibility that the nonsibilant fricatives may be identified correctly in this experiment, i.e., the listeners were able to successfully match the synthetic version to the natural sound perhaps from the cues of coarticulation or the information contained within the noise segment.

7.5.1 Mermelstein versus Natural (MvN) Test

The objective of this test was to determine whether the listeners were able to identify correctly the category of the synthetic fricative X. The listeners were presented with sequences of three sounds: natural fricative A, natural fricative B and synthetic fricative X. Synthetic X was derived using the Mermelstein area-derivation technique and either natural fricative A or natural fricative B were in the same speech category as X. The listeners were forced to choose from between A and B, the sound which they felt most resembled the fricative sound X.

A sequence of three sounds formed a single presentation, and the listeners were presented with a total of 96 presentations. In each presentation, each sound segment A, B and X had a duration of 0.25 s. Between each sound segment was 0.4 s of silence. After X was played, there followed one second of silence for the listener to make a decision (between natural fricative A and natural fricative B) and tick the appropriate box on the form provided before the next presentation of sounds was played. If the listeners were unable to make a decision, they had to guess. This is why this type of test is known as a 2AFC test.

The presentations were made using the natural and synthetic noise segments of the four voiceless fricatives in /a/ context; /f, θ, s, ʃ/. Each synthetic fricative X was tested 24 times against a pair of natural fricatives A and B, one of which was in the same category as the synthetic X. For example, the synthetic /f/ was tested against natural /f/ and natural /θ/ 8 times, against natural /f/ and natural /s/ 8 times, and against natural /f/ and natural /ʃ/ 8 times, giving a total of 24 presentations altogether for the fricative /f/. To ensure fair comparisons, the order of the (correct) natural fricatives was reversed 50% of the time, which means that the probability that synthetic X was in the same category as natural A was 50% and in the same category as natural B also 50%.

	Synthesised			
Natural	/f/	/θ/	/s/	/ʃ/
/(a)f/	**180**	75	38	27
/θ/	71	**183**	30	23
/s/	82	84	**293**	45
/ʃ/	98	90	71	**337**

TABLE 7.1: Final scores for the Mermelstein versus natural test of 18 listeners. Values along the diagonal indicate the number of times the listener identified the correct category of the synthetic fricative. Values off the diagonal indicate the number of times the listeners chose a different category than that was intended for the synthetic fricative.

	Synthesised			
Natural	/f/	/θ/	/s/	/ʃ/
/(a)f/	**227**	66	45	43
/θ/	72	**198**	56	39
/s/	69	78	**249**	43
/ʃ/	64	90	82	**307**

TABLE 7.2: Final scores for the Blum versus natural test of 18 listeners. See caption for Table 7.1.

The order and combination of the presentations that were used are given in Table B.1 in the appendix.

The scores for each subject are presented in Tables B.2–B.5 of the appendix. The final scores obtained for all the listeners in this test are presented here in Table 7.1. Entries along the diagonal indicate the number of times the listeners successfully identified the category of the synthetic fricative sound, while entries off the diagonal indicate the number of times the listeners preferred the other category of sound over the correct category. These results will be further analysed in Section 7.6 using statistical methods of interpretation.

7.5.2 Blum versus Natural (BvN) Test

The setup for the BvN test is similar to that of the MvN test described in the previous section, with the exception that the synthetic segments were derived using the Blum transform. The order and combination of the presentations that were used are given in Table B.6 in the appendix. The final scores obtained for all the listeners in this test are presented here in Table 7.2. The scores for each subject are presented in Tables B.7–B.10 of the appendix.

	Total Score	
Fricative	Mer	Blum
/f/	74	106
/θ/	96	84
/s/	103	77
/ʃ/	111	69

TABLE 7.3: Final scores for the Mermelstein (denoted Mer) versus Blum test for 18 subjects. The values indicate the number of times the listener chose the Blum or Mermelstein area-derivation technique for the fricative in question.

7.5.3 Mermelstein versus Blum (MvB) Test

The objective of this test was to determine whether the listeners preferred one area-derivation technique over the other for the synthetic fricatives. The listeners were presented with sequences of three sounds: synthetic fricative A, synthetic fricative B and natural fricative X. Sounds A, B and X were in the same speech category with sounds A and B derived using the Blum area-derivation technique and the Mermelstein area-derivation technique. The listeners were forced to choose between A and B, the sound which they felt most resembled the natural fricative sound X. The listeners were presented with a total of 40 presentations.

Each natural fricative X was tested 10 times against a pair of synthetic fricatives A and B in the same speech category. For example, the natural /f/ was tested against synthetic /f/ derived using the Mermelstein technique and synthetic /f/ derived using the Blum transform 10 times. To ensure fair comparisons, the order of the synthetic fricatives was reversed 50% of the time, as was done in the MvN and BvN tests.The order and combination of the presentations used are given in Table B.6 in the appendix.

The scores for each subject are presented in Table B.12 of the appendix. The final scores obtained for all the listeners in this test are presented in Table 7.3. The second column denotes the number of times the listeners decided that the Mermelstein-derived fricative sounded more like the natural fricative, while the third column denotes the number of times the listeners decided that the Blum-derived fricative sounded more similar to the natural fricative. These results will be further analysed in Section 7.6 using statistical methods for interpretation.

7.5.4 Branches versus No Branches (BvNB) Test

The objective of this test was to determine whether the listeners decided that the synthetic version of the fricative with inclusion of the side branches sounded more similar to the natural fricative than the synthetic version modelled without the side branches. The listeners were presented with sequences of three sounds: synthetic fricative A,

	Total Score	
Fricative	SC	NB
/ʃ/*	28	116
/(a)f/*	75	69
/(i)f/*	83	61
	SC+PS	NB
/s/$^{+}$	70	74
/ʃ/**	34	110
/(a)f/**	73	71
/(i)f/**	89	55

TABLE 7.4: Final scores for the BvNB test for 18 subjects. PS, SC and NB denote the pyriform sinuses, sublingual cavity and no branches respectively. The fricative /s/$^{+}$ was modelled only with inclusion of the pyriform sinuses as no sublingual cavity was evident in the production.

synthetic fricative B and natural fricative X. Sounds A, B and X are in the same speech category with sounds A and B derived using the Blum area-derivation technique, one with and one without inclusion of the areas of the side branches in the modelling. The listeners were forced to choose between A and B, the sound which they felt most resembled the natural fricative sound X. The listeners were presented with a total of 56 presentations.

Each of these fricative was tested eight times against a synthetic pair modelled with and without side branches. For example, the natural fricative /f/ was tested eight times against the synthetic fricative /f/ with side branches and with synthetic /f/ without side branches. As with the previous tests, the order of the synthetic branched and branchless models was reversed 50% of the time. The order and combination of the presentations used are given in Table B.13 of the appendix.

The scores for each subject are presented in Table B.14 of the appendix. The final scores obtained for all the listeners of this test are presented in Table 7.4. The second column denotes the number of times the listeners decided that the synthetic fricative with the areas of the side branches included in the model sounded more like the natural fricative, while the third column denotes the number of times the listeners decided that the synthetic fricative without side branches sounded more similar to the natural fricative.

7.5.5 Rate the Sound (RtS) Test

The objective of this test was to assess how well the synthetic segments sounded to the listeners when they were actually informed of the intended category of the sound. This test was given last; before taking this particular test the listeners had not been aware that they were listening to synthetic and natural speech sounds. At the start of this

Score	Description
1	Very poor. Does not sound like the identified sound.
2	Poor. I wouldn't know what the sound was if information were not provided.
3	Medium. I am not sure.
4	Good. This is a good version of the identified sound.
5	Very good. This sounds like the real thing.

TABLE 7.5: Rating scheme for RtS test

test, the listeners were informed (from the instructions) that they would be listening to speech sounds and during the course of the test, the listeners were informed of the identified category of each sound they were hearing. For instance, if an /f/ sound was played, the information on the form states "/f/ as in *f*un". The listeners however, were not informed if the sound that was played was a synthetic or naturally-uttered version of the fricative.

For this test, the listeners were required to rate the sound on a scale of 1 to 5, based on the rating scheme given in Table 7.5. As with the previous four tests, if they were unable to make a decision, then they had to guess within the allocated time.

Each fricative, natural and synthetic, was repeated 5 times in random order throughout the test for the listeners to rate. Since there were a total of 25 sound segments described earlier in this chapter, this will give a total of 125 sounds for the listener to rate. Each sound segment had a duration of 0.25 s. After each sound was played, there followed two seconds of silence for the listeners to make a decision by circling the appropriate number on the form provided, before the next sound presentation was played.

The order of the sound segments and information on the segments are shown in Appendix B.15. Scores given by each of the 18 listeners are given in Appendices B.16–B.18. The averaged scores of each sound segment given by the listeners are shown in Table 7.6.

Each fricative sound listed in Table 7.6 was rated 5 × 18 = 90 times. We assume the averaged values are a fair representation of the quality of the fricative sound. It is also worth noting that statistical t-tests on these values show that they are statistically significant, only if the difference in the values is greater than 1. The following observations are based on the values of Table 7.6.

1. The nonsibilant fricatives received the lowest ratings, ranging from 1.8 to 2.3. The lowest rating of 1.8 was received by the Mermelstein-derived /(i)f/ and the Blum-derived /θ/. The highest rating of 2.3 was received by the natural /(u)f/.

2. The sibilants had the highest ratings, ranging from 2.5 to 3.9. For each sibilant fricative, the Blum-derived sounds received the highest rating of 3.0 and 3.9 for the fricative sounds /s/ and /ʃ/ respectively.

	Total Score			
Fricative	Natural	Mer	Blum	Blum*
/(a)f/	2.0	2.0	2.0	-
/(a)f/*	-	-	-	2.1
/(a)f/**	-	-	-	2.0
/(i)f/	2.1	1.8	1.9	-
/(i)f/*	-	-	-	1.8
/(i)f/**	-	-	-	1.8
/(u)f/	2.3	2.2	1.9	-
/θ/	1.9	2.2	1.8	-
/s/	2.5	2.9	3.0	-
/s/$^+$	-	-	-	2.9
/ʃ/	3.5	3.8	3.9	-
/ʃ/*	-	-	-	3.5
/ʃ/**	-	-	-	3.8

TABLE 7.6: Final results of the RtS test. The total scores of 18 listeners were averaged to match the scale of the rating scheme shown previously in Table 7.5. Mer denotes Mermelstein and Blum* denotes the Blum models with inclusion of the side branches respectively. The sign $^+$ denotes inclusion of the pyriform sinuses in the modelling, * for inclusion of the sublingual cavity and ** for inclusion of both.

3. Addition of the side branches did not result in higher scores compared to versions derived without the side branches.

4. In some cases, the synthetic sounds received a higher rating than the natural sounds. It is possible that the effects of down-sampling and windowing on the recorded signal may have distorted the quality of the sound, making them comparable to the synthetic versions, although no actual experiment was conducted to confirm this.

The Blum-derived sounds modelled without the side branches seem to have the highest scores for the sibilant fricatives. Surprisingly, the Mermelstein-derived sounds received an overall higher rating for the nonsibilant fricatives. This is unexpected, since the previous analysis between the measured and predicted spectra had shown that the Blum-derived spectra provided better matches in terms of gross spectral shape, particularly in the case of /θ/ and /(a)f/.

7.6 Statistical Analysis

A statistical analysis of the results of the first four listening experiments was made in order to determine whether there is sufficient evidence to draw solid conclusions from the data, the conclusions here depending on the type of test. In general, the aim was to determine whether the listeners picked the correct category of sound because they found the synthetic sound very much alike to the natural sound, or whether it was just a lucky guess. The listeners, after all, did not know they were actually comparing speech sounds (in the first four tests). If it is possible to determine that the decisions of the listeners were significantly affected by the category of sound (i.e., type of fricative) or area-derivation technique, then these findings can be used to speculate on the important features of the spectra, the important articulatory properties and the importance of the pyriform sinuses and sublingual cavity.

Several terms will be defined here for clarity. The *population* of this analysis is the sound segments. A *sample* is a part of the population, e.g., the sound segments derived using the Mermelstein technique. One *datum* is the score for a sound segment obtained by one listener. For the BvN, MvN, BvNB and MvB tests, this is the total number of times the listener chose a particular sound segment. *Data* is a set of values (i.e., scores from all the listeners) which correspond to a sample. Finally and most importantly, the *level of significance* is denoted by α and fixed at 0.05. If the computed probability, known as the P-value, of obtaining the observed difference—assuming the null hypothesis—is less than the level of significance, then there is sufficient evidence to reject the null hypothesis[1]. This implies that the difference in the mean of the sample data is statistically significant and does not result from chance and/or sampling errors.

The following sections describe the two statistical tools that were used to process the results of the listening tests. Examples will be given of how the statistical tools were applied specifically to the data. We follow this with the results obtained from the analysis. The reader can refer to Johnson (1988) for more information on statistical analysis.

7.6.1 Application of t-Test

One of the two statistical tools adopted for this investigation is the t-test which is commonly used for inferences involving two groups of data. The t-test determines whether there is a significant difference between the observed means of the sample data. The samples that are used in this analysis are classified as *dependent* samples, because comparisons are made of each datum (i.e., score) of the first sample with the datum in the second sample, both of which derive from the same source (i.e., same listener). The two data from each sample set (of the same source) are known as paired data, and the numerical difference between them is known as the paired difference, denoted by d. The distribution of the paired difference (the t distribution) is assumed to be normal with a mean value of μ_d and a standard deviation of σ_d. In effect, the mean difference of the two dependent populations is tested using the observed means of the paired differences. The final computed value of the t-test is denoted as t.

[1] Specifically, if $0.01 < P < 0.05$ then there is some evidence to reject H_0. If $0.001 < P < 0.01$ then there is strong evidence to reject H_0. And if $P < 0.001$ then there is very strong evidence to reject H_0

No.	Subject	Total Mer (M)	Total Blum (B)	d (M−B)	d^2
1	ZH	7	3	4	16
2	AT	8	2	6	36
3	SC	2	8	−6	36
4	BC	4	6	−2	4
5	SB	5	5	0	0
6	DL	5	5	0	0
7	CB	1	9	−8	64
8	JG	8	2	6	36
9	SP	4	6	−2	4
10	BN	7	3	4	16
11	MN	7	3	4	16
12	BL	4	6	−2	4
13	JD	7	3	4	16
14	SS	3	7	−4	16
15	JP	5	5	0	0
16	AC	5	5	0	0
17	ES	7	3	4	16
18	AH	7	3	4	16
Total				12	296

TABLE 7.7: The results of the MvB test for synthetic /θ/ for 18 listeners. Total (M) and Total (B) denotes the number of times the listener chose the Mermelstein- (denoted Mer) or the Blum-derived fricative respectively out of a total of 10 times.

An example is given here for clarity. We wish to determine whether the scores of the synthetic /θ/ of the MvB test were statistically significant, that is, if the type of area-derivation technique affected the choices made by the listeners. The null hypothesis states:

$$H_0(/\theta/) : \mu_M = \mu_B, \tag{7.4}$$

that is, the mean, μ_M and μ_B of the scores of the Mermelstein-derived /θ/ and the Blum-derived /θ/, respectively, are equal, and the area-derivation technique did not influence the choices made by the listeners. The data used for the t-test are the total scores obtained by the 18 listeners shown in Table 7.7. The paired difference, d, shown in the fifth column was computed by finding the difference between the paired scores of the Mermelstein and Blum techniques. The last column shows the squared value of this difference, denoted d^2.

The observed mean value of d is given by:

$$\bar{d} = \frac{\sum d}{n} \tag{7.5}$$

where n is the number of listeners. From the above equation, the computed $\bar{d}$ for this example is:

$$\begin{aligned}\bar{d} &= \frac{12}{18} \\ &= 0.666\end{aligned}$$

The standard deviation of the data is given by:

$$\begin{aligned}s_d &= \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} \\ &= \sqrt{\frac{296 - \frac{(12)^2}{18}}{17}} \\ &= 4.116\end{aligned} \tag{7.6}$$

The inferences are completed using the t distribution:

$$\begin{aligned}t &= \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \\ &= \frac{0.666 - 0}{4.116/\sqrt{18}} \\ t &= 0.687\end{aligned} \tag{7.7}$$

The degrees of freedom are computed from $df = n - 1$, with n equal to the number of listeners (i.e., 18). The t-test used here is two-tailed (both sides of the distribution curve are used to represent the critical region); thus the level of significance is halved (i.e., 0.025) when finding the critical values. By looking at a table of the t distribution, the critical values for the given example are $\pm t(17, 0.025) = \pm 2.11$. If the calculated value of t falls in the critical region, then the null hypothesis is rejected and we conclude that the alternative hypothesis is correct. Since $t = 0.687$ does not fall in the critical region, then in this case we fail to reject the null hypothesis.

The area at the ends of the t distribution curve up to the t-value represent the probability, known as the P-value, of the null hypothesis. Alternatively, if P is below the level of significance of $\alpha = 0.05$, then there is sufficient evidence to reject the null hypothesis. For the example given above, the computed probability of t, assuming the null hypothesis is $P = 0.501$, which is higher than α. As previously stated, we fail to reject the null hypothesis. Therefore, for the fricative /θ/, the area-derivation technique did not influence the decisions made by the listeners.

7.6.2 Application of Analysis of Variance (ANOVA)

The ANOVA test is a test of a hypothesis about the means of three or more groups of data. If we have four groups of data, one for each fricative, the null hypothesis is represented by:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4, \tag{7.8}$$

where μ_1, μ_2, μ_3 and μ_4 are the mean values of the first, second, third and fourth groups of data respectively. Application of the t-test will not be convenient in this case, as this will require additional work in order to test the null hypothesis;

$$\begin{aligned} H_1 &: \mu_1 = \mu_2, \\ H_2 &: \mu_1 = \mu_3, \\ H_3 &: \mu_1 = \mu_4, \\ H_4 &: \mu_2 = \mu_3, \\ H_5 &: \mu_2 = \mu_4, \\ H_6 &: \mu_3 = \mu_4. \end{aligned}$$

Rejection of any of the test hypotheses of the two means (shown above) is sufficient to reject the null hypothesis that all four means are equal. However, it is evident that as the number of groups increases, so will the number of test pairs. Application of ANOVA allows a single test of the null hypothesis (that all means are equal) against the alternative hypothesis (at least one is different) with a specified value of α.

For clarity, an example of an ANOVA test for the /s/ scores of the MvN test is described here. Table 7.8 shows the number of times listeners chose the natural sounds /f/, /θ, /s/, and /ʃ/ when compared against the synthesised /s/. The number of times each listener correctly selected natural /s/ out of 24 presentations is shown in column 3. However, one must remember that the presentations of the synthetic /s/ were always compared with the natural /s/ paired with one of three other natural fricatives (i.e., /f/, /θ/ or /ʃ/). This means that the probability that a listener would choose the correct category of sound is actually three times more than choosing an incorrect one. Because of this, the scores in column 3 were divided by 3 and the new values are shown in column 4. Thus the maximum possible value of column 4 is now reduced to 8, to match the highest possible value in the other remaining columns. The values in the last four columns are the ones that are used for the ANOVA test.

The null hypothesis for this test is given by:

$$H_0(/\mathrm{s}/) : \mu_\mathrm{f} = \mu_\theta = \mu_\mathrm{s} = \mu_\int, \tag{7.9}$$

No.	Subject	Total	/s/*	/f/	/θ/	/ʃ/
1	ZH	21	7.00	1	1	1
2	AT	18	6.00	2	1	3
3	SC	18	6.00	2	0	4
4	BC	24	8.00	0	0	0
5	SB	12	4.00	1	3	8
6	DL	15	5.00	4	3	2
7	CB	13	4.33	5	2	4
8	JG	10	3.33	4	2	8
9	SP	17	5.67	0	1	6
10	BN	19	6.33	2	3	0
11	MN	21	7.00	0	0	3
12	BL	14	4.67	3	0	7
13	JD	14	4.67	3	1	6
14	SS	16	5.33	2	1	5
15	JP	12	4.00	3	6	3
16	AC	15	5.00	3	2	4
17	ES	13	4.33	3	4	4
18	AH	21	7.00	0	0	3
Column Total			$C_1 = 97.67$	$C_2 = 38$	$C_3 = 30$	$C_4 = 71$
Column Data Size			$k_1 = 18$	$k_2 = 18$	$k_3 = 18$	$k_4 = 18$

TABLE 7.8: The results of the MvN test for synthetic /s/ for 18 listeners. Total denotes the number of times each listener correctly chose /s/ over /f, θ, ʃ/ out of 24 times. In /s/* the total was divided by three. The last four columns were the data used for ANOVA.

where μ_{f}, μ_{θ}, μ_{s} and $\mu_{\int}$ are the mean values for the scores of each fricative category. That is, the category of the synthetic sound /s/ was not identified by the listeners and did not affect the decisions that were made. The alternative to the null hypothesis is:

$$H_a(/\text{s}/): \quad \begin{array}{c}\text{the identity of the synthetic sound /s/ has a} \\ \text{significant effect on the decisions made by the listener.}\end{array} \tag{7.10}$$

Thus, the null hypothesis can be rejected if one or more of the means are significantly different from the others. The decision to reject H_0 or fail to reject H_0 is made using the F distribution and the F test statistic (do not be confused with the non-italicised F which represents "fricative"). As we shall see from the example, computation of F is in effect a ratio of two variances.

First, the total sum of squares, SS(total), of the data is computed:

$$\text{SS(total)} = \sum(x^2) - \frac{(\sum x)^2}{n} \tag{7.11}$$

where x represents one datum and n the total sample size respectively. Here, x are the values in the last four columns of Table 7.8 and $n = \sum_{i=1}^{4} k_i = 72$. Thus from

Equation 7.11:

$$\begin{aligned}
\sum(x^2) &= 1152.89 \\
\sum x &= 236.67 \\
\mathrm{SS(total)} &= 1152.89 - \frac{236.67^2}{72} \\
&= 374.95
\end{aligned}$$

Next, SS(total) is divided into two parts: the sum of squares due to the column factor, SS(factor), and the sum of squares due to experimental error, SS(error). This is called partitioning as SS(factor) + SS(error) = SS(total). The sum of squares due to the factor—SS(fricative) in this case—measures the variation between the factor levels (i.e., fricative scores) and is computed from:

$$\mathrm{SS(factor)} = \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \ldots\right) - \frac{(\sum x)^2}{n}, \tag{7.12}$$

where C_i and k_i is the column total and sample size respectively and is given in Table 7.8. Thus from Equation 7.12, the variation due to the fricative, SS(fricative), is:

$$\begin{aligned}
\mathrm{SS(fricative)} &= \left(\frac{97.67^2}{18} + \frac{38^2}{18} + \frac{30^2}{18} + \frac{71^2}{18}\right) - \frac{236.67^2}{72}, \\
&= 162.28
\end{aligned}$$

The sum of squares, SS(error), measures the variations within the rows and is given by:

$$\mathrm{SS(error)} = \sum(x^2) - \left(\frac{C_1^2}{k_1} + \frac{C_2^2}{k_2} + \frac{C_3^2}{k_3} + \ldots\right) \tag{7.13}$$

Thus from Equation 7.13 the variations in the rows of Table 7.8 is:

$$\begin{aligned}
\mathrm{SS(error)} &= 1152.89 - \left(\frac{97.67^2}{18} + \frac{38^2}{18} + \frac{30^2}{18} + \frac{71^2}{18}\right) \\
&= 212.68
\end{aligned}$$

The degrees of freedom, *df*, associated with the three sources are determined as follows:

$$df(\mathrm{factor}) = c - 1 \tag{7.14}$$

$$df(\mathrm{total}) = n - 1 \tag{7.15}$$

$$df(\mathrm{error}) = (k_1 - 1) + (k_2 - 1) + (k_3 - 1) + \ldots \tag{7.16}$$

where c is the number of samples (i.e., number of columns), n is the total number of data and k_i is the number of data in each sample. For the example given here, the computed *df*'s are:

$$\begin{aligned} df(\text{factor}) &= 3 \\ df(\text{total}) &= 71 \\ df(\text{error}) &= 68 \end{aligned}$$

The mean squares, MS, of the sources are obtained by dividing the sum of squares with the corresponding number of degrees of freedom:

$$\text{MS(factor)} = \frac{\text{SS(factor)}}{\text{df(factor)}} \tag{7.17}$$

$$\text{MS(error)} = \frac{\text{SS(error)}}{df(\text{error})} \tag{7.18}$$

Thus from the equations above, the MS(factor) and MS(error) are:

$$\begin{aligned} \text{MS(factor)} &= \frac{162.28}{3} \\ &= 54.09 \\ \text{MS(total)} &= \frac{212.68}{68} \\ &= 3.13 \end{aligned}$$

The hypothesis is now complete with the two mean squares as measures of variance. The calculated value of the test statistic, F, is found by:

$$\begin{aligned} F &= \frac{\text{MS(factor)}}{\text{MS(error)}} \\ &= \frac{54.09}{3.13} \\ F &= 17.29 \end{aligned} \tag{7.19}$$

The decision to reject H_0 or fail to reject H_0 is made by comparing the calculated value of $F = 17.29$ to a one-tailed critical value of the F distribution (given in tables in most statistics books). If the value of F falls in the critical region, then the null hypothesis can be rejected. As stated earlier, the level of significance, α, was settled at 0.05, and if the probability, P, of F, assuming the null hypothesis, is less than α, then there is sufficient evidence to reject the null hypothesis. In this example, $P < 0.0001$, thus there is sufficient evidence to reject the null hypothesis and we can conclude that the listeners had made their decisions based on identification of the sound. Whether or not the decisions were correct will be discussed in depth in the following sections.

7.6.3 ANOVA: Results of MvN and BvN Tests

The ANOVA test was applied to the results of the MvN and BvN tests because in each case the hypothesis regarding the observed means of four groups of data were being tested. The data that were used for ANOVA are the number of times the listener selected the category of each natural fricative (i.e., /f, θ, s, ʃ/), given the synthetic fricative X. An example of ANOVA on the results for synthetic /s/ of the MvN test was given in the previous section. Statistical analysis of the test results are presented here in one section, because, apart from the area-derivation technique, the two tests were exactly the same.

The objective of tests MvN and BvN was to determine whether the listeners were able to identify correctly the category of the synthetic fricative. The results of these two tests were already presented in Tables 7.1 and 7.2 (see page 125) for the MvN and BvN tests, respectively. The values along the diagonal indicate the number of times the listeners chose the correct fricative category, while values off the diagonal indicate the number of times the listener chose a different category of the intended sound. As shown in the example given in the previous section, the values along the diagonal were divided by three to match the maximum scores of the remaining columns. The ANOVA test was applied on each column of scores to determine whether the listener significantly chose the correct category of sound based on identification of the fricative.

The results of the ANOVA test are shown in Tables 7.9 and 7.10 for the MvN and BvN test, respectively. The second column lists the means of each data sample (i.e., average of the number of times the natural fricative was selected), so that the results can be interpreted correctly. Only the computed F value and the probability of the computed F, denoted P, which assumes the null hypothesis, is given, since these are the values which determine whether there is sufficient evidence to reject the null hypothesis.

The results of ANOVA have allowed the following interpretations from the MvN test results.

1. For synthetic /f/, the listeners had made their decisions based on the sound but they identified the sound incorrectly as category /ʃ/ or /s/ most of the time, shown by the higher mean values for /ʃ/ and /s/ compared to /f/.
2. For synthetic /θ/, the listeners did not make decisions based on the sound. They did not recognise the sound and thus, decisions were made at random.
3. For synthetic /s/ and /ʃ/, the listeners had made their decisions based on identification of the sounds and significantly chose the correct category of sound.

The results of ANOVA have allowed the following interpretations from the BvN test results.

Fricative	Mean	F	P	Conclusion
/f/	$\mu_{[f]} = 3.33$ $\mu_{[\theta]} = 3.94$ $\mu_{[s]} = 4.56$ $\mu_{[\int]} = 5.44$	5.337	0.0023	Reject H_0
/θ/	$\mu_{[f]} = 4.72$ $\mu_{[\theta]} = 3.39$ $\mu_{[s]} = 4.67$ $\mu_{[\int]} = 5.00$	2.033	0.12	Fail to reject H_0
/s/	$\mu_{[f]} = 2.11$ $\mu_{[\theta]} = 1.67$ $\mu_{[s]} = 5.43$ $\mu_{[\int]} = 3.94$	17.29	< 0.0001	Reject H_0
/ʃ/	$\mu_{[f]} = 1.50$ $\mu_{[\theta]} = 1.28$ $\mu_{[s]} = 2.50$ $\mu_{[\int]} = 6.20$	63.44	< 0.0001	Reject H_0

TABLE 7.9: ANOVA results of synthetic /f, θ, s, ʃ/ for MvN test. The level of significance is $\alpha = 0.05$. Therefore, if the computed P is lower than this value there is sufficient evidence to reject the null hypothesis (denoted H_0) and the listeners had made the decisions based on recognition of the sound.

Synthetic	Mean	F	P	Conclusion
/f/	$\mu_{[f]} = 4.20$ $\mu_{[\theta]} = 4.00$ $\mu_{[s]} = 3.83$ $\mu_{[\int]} = 3.56$	0.9319	0.43	Fail to reject H_0
/θ/	$\mu_{[f]} = 4.22$ $\mu_{[\theta]} = 3.67$ $\mu_{[s]} = 4.33$ $\mu_{[\int]} = 5.00$	1.530	0.21	Fail to reject H_0
/s/	$\mu_{[f]} = 2.50$ $\mu_{[\theta]} = 3.11$ $\mu_{[s]} = 4.61$ $\mu_{[\int]} = 4.56$	7.118	0.0003	Reject H_0
/ʃ/	$\mu_{[f]} = 2.39$ $\mu_{[\theta]} = 2.17$ $\mu_{[s]} = 2.39$ $\mu_{[\int]} = 5.69$	23.27	< 0.0001	Reject H_0

TABLE 7.10: ANOVA results of synthetic /f, θ, s, ʃ/ for BvN test. The symbol μ denotes the mean of the scores for each tested fricative. The level of significance is $\alpha = 0.05$. Therefore, if the computed P is lower than this value there is sufficient evidence to reject the null hypothesis (denoted H_0) and the listeners had made their decisions based on recognition of the sound.

1. For synthetic /f/ and /θ/, the listeners did not make decisions based on identification of the sound. They did not recognise the sounds and thus, decisions were made at random.

2. For /s/ the listeners made their decisions based on the sound, but they mistakenly identified /s/ as /ʃ/ almost the same number of times as they correctly identified it as /s/, resulting in similar mean values for /s/ and /ʃ/.

3. For synthetic /ʃ/ the listeners had made their decisions based on identification of the sound and significantly chose the correct category of sound.

The results of this test suggests that both the Mermelstein and Blum area-derivation techniques failed to produce synthetic sounds for the nonsibilant fricatives /f/ and /θ/ that are identifiable. ANOVA on the MvN test results shows that the listeners mistakenly identified /f/ as /s/ or /ʃ/ most of the time while /θ/ was not identified at all. For the BvN test results, both /f/ and /θ/ failed to be identified by the listeners. For the Blum-derived sounds this is quite a disappointment since comparison between the predicted and measured-counterpart spectra showed a good match.

For the sibilant fricatives, the Mermelstein area-derivation technique produced a synthetic version that was better than the Blum area-derivation technique as the listeners significantly identified the sound correctly. The Blum-derived /s/ seems to be mistakenly identified as /ʃ/ as much as it was identified correctly as /s/. For /ʃ/, both area-derivation techniques produced synthetic versions that were good enough to be identified correctly by the listeners. By looking at the mean values for the synthetic /ʃ/ from both listening tests, the higher mean for the MvN test suggests that the Mermelstein-derived sound was more identifiable than the Blum counterpart.

7.6.4 *t*-Test: Results of MvB Test

The *t*-test was applied to the results of the MvB test because the hypothesis of the observed means of two groups of data is being tested. Recall that the objective of the MvB test was to determine the area-derivation technique which produces a synthetic version of the fricative that sounds most like the natural one. The results of this test were already presented in Table 7.3. The objective of the *t*-test was therefore to determine whether the decisions that were made by the listeners were influenced by the area-derivation technique or by chance or sampling errors. The data that were used for the *t*-test are the number of times the listener selected the Mermelstein-derived sound and the Blum-derived sound given the natural fricative X. An example of application of the *t*-test for /θ/ has already been shown in Subsection 7.6.1. The critical values are $\pm t(17, 0.025) = \pm 2.11$ for 17 degrees of freedom and $\alpha = 0.05$ level of significance. Each paired difference, d, was computed by subtracting the Blum datum from the Mermelstein

Fricative	Mean	t	P	Conclusion
/f/	$\mu_M = 4.11$ $\mu_B = 5.89$	-2.12	0.050	Reject H_0
/θ/	$\mu_M = 5.33$ $\mu_B = 4.67$	0.687	0.501	Fail to reject H_0
/s/	$\mu_M = 5.72$ $\mu_B = 4.28$	1.56	0.137	Fail to reject H_0
/ʃ/	$\mu_M = 6.17$ $\mu_B = 3.83$	2.12	0.050	Reject H_0

TABLE 7.11: The t-test analysis on the MvB test results. The symbols μ_M and μ_B denote the mean of the scores of the Mermelstein-derived and the Blum-derived sounds respectively. The level of significance is $\alpha = 0.05$. Therefore, if the computed P is lower than this value there is sufficient evidence to reject the null hypothesis (denoted H_0) and the listeners had made their decisions based on the area-derivation technique.

datum. Therefore, if t is negative, the Blum area-derivation technique was preferred by the listener and if t is positive the Mermelstein area-derivation technique was preferred. The results of the t-test are shown in Table 7.11.

The t-test allowed the following interpretations from the MvB test results.

1. For /f/, the listeners significantly decided that the Blum-derived version sounded more like the natural sound.
2. For /θ/ and /s/, the listeners decided that both area-derivation techniques produced synthetic versions that did not sound like the natural sound.
3. For /ʃ/, the listeners significantly decided that the Mermelstein-derived version sounded more like the natural sound.

These results do not contradict the conclusions previously drawn from the MvN and BvN tests. The nonsibilant /f/ was mistakenly identified as a sibilant fricative in the MvN test and was not identified at all in the BvN test. The MvB test results have shown however, that the Blum-derived version sounded more similar to the natural /f/ compared to the Mermelstein-derived version.

The fricative /θ/ was unidentified in both the MvN and BvN tests. The MvB test shows that both area-derivation techniques failed to produce synthetic versions that sounded similar to the natural sound. This is quite disappointing for the Blum-derived /θ/, whose predicted spectrum was well-matched to the measured spectrum.

For /s/, the listeners were able to identify the correct category of sound for the MvN test. For the BvN test, the listeners were able to identify correctly the sound as being a sibilant fricative (i.e., /s/ or /ʃ/). The MvB test results show that the area-derivation techniques did not significantly influence the decisions made by the listeners. However, the Mermelstein-derived sound does have a higher mean value, in agreement with the results of the MvN and BvN tests.

Finally, for /ʃ/ the MvN and BvN tests have shown that the listeners were able to identify the fricative correctly. The MvB results show that the listeners significantly found the Mermelstein-derived version more alike to the natural sound. This is also in agreement with the mean values previously computed for ANOVA of /ʃ/, where the mean of the MvN test was slightly higher than that of the BvN test.

7.6.5 *t*-Test: Results of BvNB Test

The *t*-test was also applied to the results of the BvNB test. Recall that the objective of this test was to determine whether inclusion of the side branches produced a synthetic version that sounded more like the natural one. The results of this test were already presented in Table 7.4. The objective of the *t*-test was therefore to determine whether the decisions made by the listeners were influenced by inclusion of the side branches or entirely by chance or sampling errors. The data that were used for the *t*-test are the number of times the listener selected the branchless version of sound and the branched version of sound given the natural fricative X. Application of the *t*-test to the data is similar to the example shown in Subsection 7.6.1. The critical values are $\pm t(17, 0.025) = \pm 2.11$ for 17 degrees of freedom and $\alpha = 0.05$ level of significance. Each paired difference, d, was computed by subtracting the branched datum from the branchless datum. Therefore if t is positive the synthetic sound with inclusion of the side branches was preferred by the listener and if t is negative the branchless version was preferred.

The synthetic /s/ was derived with the pyriform sinuses only as the sublingual cavity was not evident during production. The synthetic /ʃ/, /(a)f/ and /(i)f/ were first derived with the sublingual cavity only, and then with both sublingual cavity and pyriform sinuses. All the synthetic sounds here were derived using the Blum area-derivation technique. The results of the *t*-test are shown in Table 7.12.

The *t*-test allowed the following interpretations from the BvNB test results:

1. For the fricatives /s/, /(a)f/ and /(i)f/, addition of the side branches in the derivation of the synthetic sound did not influence the decisions made by the listeners.
2. For the fricative /ʃ/, the listeners significantly decided that exclusion of the side branches produced a synthetic version that sounded more like the natural sound.

These results clarify the findings of Chapter 5, where it was found that addition of the side branches in the model did not improve the degree of match between the predicted

Fricative	Mean	t	P	Conclusion
/s/$^+$	$\mu_{\text{B}} = 3.89$ $\mu_{\text{NB}} = 4.11$	−0.301	0.767	Fail to reject H_0
/ʃ/*/	$\mu_{\text{B}} = 1.56$ $\mu_{\text{NB}} = 6.44$	−8.02	0.000	Reject H_0
/ʃ/**	$\mu_{\text{B}} = 1.89$ $\mu_{\text{NB}} = 6.11$	−5.13	0.000	Reject H_0
/(a)f/*	$\mu_{\text{B}} = 4.17$ $\mu_{\text{NB}} = 3.83$	0.566	0.579	Fail to reject H_0
/(a)f/**	$\mu_{\text{B}} = 4.06$ $\mu_{\text{NB}} = 3.94$	0.148	0.884	Fail to reject H_0
/(i)f/*	$\mu_{\text{B}} = 4.61$ $\mu_{\text{NB}} = 3.39$	1.42	0.172	Fail to reject H_0
/(i)f/**	$\mu_{\text{B}} = 4.94$ $\mu_{\text{NB}} = 3.06$	1.99	0.063	Fail to reject H_0

TABLE 7.12: The t-test analysis on the BvNB test results. $^+$ denotes inclusion of the pyriform sinuses in the modelling, * for inclusion of the sublingual cavity and ** for inclusion of both. The symbols μ_{B} and μ_{NB} denote the mean of the scores of models derived with and without the inclusion of side branches respectively. The level of significance is $\alpha = 0.05$. Therefore, if the computed P is lower than this value there is sufficient evidence to reject the null hypothesis (denoted H_0)and the listeners had made their decisions based on inclusion/exclusion of the side branches in the modelling.

and measured spectra for all the fricatives (refer back to Table 6.2, page 111). It does seem here that, particularly for /ʃ/, the listeners decided that the sound was significantly better *without* side branches, consistent with the previous observation that the degree of match was reduced by approximately one third when side branches (sublingual cavity alone, and both sublingual cavity and pyriform sinuses) were added to the model. In the discussion of Chapter 6, we had suggested on the possibility that inclusion of side branches might not make an improvement to the modelling. The finding here seems to support this.

7.7 Perceptually Important Features of Fricative Spectra

This section discusses the results of the listening experiments and compares them with the results of the qualitative analysis presented in Chapter 6. The discussion will be divided into three main sections: sibilant fricatives, nonsibilant fricatives and side branches. Note that the term "degree of match" refers to the D_m value of the previous quantification analysis in Section 6.4.

7.7.1 Sibilant Fricatives

ANOVA analysis of the MvN and BvN tests have revealed at the 95% confidence level that for /s/ and /ʃ/ the listeners made their decisions based on identification of the sound. The Mermelstein-derived versions of both /s/ and /ʃ/ were better than the Blum-derived versions. For /s/, the listeners significantly chose the correct category of sound (i.e., natural /s/) but the Blum-derived /s/ was incorrectly identified as /ʃ/ almost half of the time. For /ʃ/, both techniques produced versions that were significantly correctly identified by the listeners but the Mermelstein-derived version received a higher mean score than the Blum-derived version, implying that it was a better match to the natural sound. Analysis of the MvB test results confirmed this; the listeners significantly chose the Mermelstein-derived /ʃ/ over the Blum-derived version as sounding more similar to the natural sound.

These observations are consistent with the findings of the qualitative analysis which previously showed that the Mermelstein-derived sibilants have higher degrees of match to the measured spectra than the Blum-derived sibilants (79% and 67% for the Mermelstein- and Blum-derived versions, respectively). These findings suggest first and foremost that the Mermelstein area derivation technique is the best area derivation technique for the sibilant fricatives. But what did the listeners think of the synthetic sounds when heard individually, without comparison to the natural counterpart? The scores for the Blum-derived and Mermelstein-derived /s/ were 3.0 and 2.9, respectively while the scores for the Blum-derived and Mermelstein-derived /ʃ/ were 3.9 and 3.8 respectively. Blum shows the higher scores overall, but the difference between the scores of each respective fricative is insignificant (separate *t*-tests have shown that for the RtS test, the difference is only significant if it is > 1). Therefore we conclude that the listeners find the synthetic—and natural—sounds on the same level of quality regardless of the area-derivation technique. Note that the natural /s/ and /ʃ/ sounds received slightly lower ratings than the synthetic sounds with 2.5 for /s/ and 3.5 for /ʃ/ (refer to Table 7.6).

If each sibilant fricative were given similar ratings, then why did the listeners decide that the Mermelstein-derived version of the sibilant fricative sounded more like the natural version than the Blum-derived version? To answer this question requires reference back to the area functions. For /ʃ/, it can now confidently be said that accurate measurements of the supra-laryngeal region are not important. The Mermelstein-derived area function had very low area measurements in this region whilst the Blum-derived area showed higher and more consistent measurements (i.e., consistent to the other area functions derived using Blum and to some other area functions found in the literature, e.g., Narayanan and Alwan 2000 (refer to Figure 4.12, page 72). But this major difference between the Blum- and Mermelstein-derived area functions resulted in very little difference between the predicted spectra; for the Blum-derived spectrum, there is a minor perturbation at 1 kHz which was absent in the Mermelstein-derived spectrum and the significant peak at 2.4 kHz was approximately 200 Hz lower in frequency and amplitude from the one in the measured- and Mermelstein-derived spectrum. The difference in the location and amplitude of the significant peak of the predicted spectra is possibly related to the differences in the effective length of the area functions; Blum measured a longer distance overall, because the grid planes were placed on an irregular longitudinal axis, while Mermelstein placed the grid planes at regular intervals along a *straight* axis. Thus, there is only one major feature of the /ʃ/ spectra which might have influenced the decision made by the listeners on the best-matched synthetic sound, and that is the significant peak located at 2.7 kHz. It is highly likely that the match in the peak location of the Mermelstein-derived spectrum with the significant peak in the measured spectrum gave the two sounds the same pitch quality mentioned by most listeners. This distinctive peak is lowered in the Blum-derived spectra, but, irrespective of this, the Blum-derived sound still sounded like an acceptable version of /ʃ/, and thus was correctly identified in the BvN test and received similar ratings as the Mermelstein- and naturally derived /ʃ/ sounds in the RtS test.

Following these observations, a hypothesis regarding the acoustic properties of /ʃ/ is presented.

1. The supra-laryngeal region is not important, or more specifically, does not need to be accurately modelled for /ʃ/. The resonances of the back cavity are found in the lower frequency region (below 2.5 kHz for subject CS) and this region does not carry the cues for the fricative's identification.

2. The precise location of the distinct peak about 2.7 kHz does not carry the cue to the fricative's identification but affects the pitch quality of the sound. A musically-trained ear is sensitive to the location of this peak and, depending on its frequency, will recognise the sound as having a pitch quality. However, even if this peak is shifted in frequency by as much as 200 Hz, the sound will still be identified as a /ʃ/ sound.

3. The high amplitude levels in the high frequency region (~3 kHz for subject CS) carry important cues to the fricative's identity. The predicted spectra imply that even an approximation of the natural energy levels may be able to produce a "good" version of /ʃ/ (with "good" based on the rating given by the listeners).

For /s/, we cannot confidently say at this point that measurements of the supra-laryngeal region are not important, because unlike /ʃ/, both area functions show similar measurements in this region. We need to be more careful regarding the interpretation of /s/; the listeners significantly identified the Mermelstein-derived /s/ sound as /s/, but the Blum-derived /s/ was sometimes confused with /ʃ/. Yet, in the MvB test, why was the area-derivation technique not significant to the listeners? One would expect the results of the MvB test to lean more towards the Mermelstein technique, since the Mermelstein-derived /s/ sound was not confused with /ʃ/ as in the Blum-derived sound.

We know from the RtS scores that the listeners decided both synthetic /s/ (and measured /s/) were relatively equal versions of the sound, which means that the differences in the spectra were irrelevant to the listener (at least up to 8.0 kHz). So there is a suspicion that the pitch quality must again have played a part in this matter. The sharp peak located at 2.8 kHz in the measured spectrum of /s/ was correctly predicted in the Mermelstein-derived spectrum, allowing the listeners to identify the sound correctly as /s/ when given the option between natural /s/ and natural /ʃ/ in the MvN test. However, the Blum-derived /s/ predicted the same peak at a lower frequency of approximately 2.7 kHz, which coincidentally is the same frequency as the significant peak in /ʃ/. Now, how will the listeners choose when presented between the natural /s/ and /ʃ/? They can either match the synthetic /s/ according to the identity of the sound (i.e., pick natural /s/) *or* according to the pitch quality of the sound (i.e., pick natural /ʃ/). At least 9 out of the 18 listening subjects had strong backgrounds in music and were very sensitive to the pitch quality of the sound. In fact, some had voluntarily mentioned that since they did not know what the sounds were, they had made their decisions based on the pitch quality alone. This is one possible explanation for finding similar mean values for the /s/ and /ʃ/ scores of the BvN test (refer to Table 7.10). This line of reasoning can also be used to explain the mean values for the Mermelstein-derived and Blum-derived /s/ of the MvB test (the mean values for Mermelstein are higher, although not significantly); when the two /s/ sounds were heard together, the cues to the identity of the fricative were much stronger and listeners less sensitive to pitch quality would not be able to tell the difference. Consequently, these results have led to the conclusion that the Mermelstein-derived /s/ is better, because the sound was not incorrectly identified as /ʃ/ in the MvN test.

We now turn to discuss the effects of including side branches into the models of the sibilant fricatives. The results of the BvNB test have revealed that their addition was only significant to listeners for the synthetic /ʃ/, who even then preferred the version of sound modelled without side cavities. For /s/, the listeners did not perceive any significant changes to the sound when the areas of the pyriform sinuses were included in the model. Again, this is in agreement with the findings of the quantitative analysis: addition of the side branches for /ʃ/ reduced the degree of match (between the predicted and measured spectra) by almost a third, while for /s/, the degree of match did not improve significantly (less than 2% increase).

The results of the listening experiment not only support the earlier suggestion that side branches do not play an important role for fricative production, but also allow several other important conclusions for /s/. It was observed that the addition of the pyriform sinuses for /s/ introduced an additional peak at about 6.0 kHz. The fact that the listeners did not perceive this change implies that it is the amplitude levels in this region that are more important to the listeners than the peak locations themselves. If this is the case, it explains why the choice of area-derivation technique for /s/ was insignificant to the listener since the major differences between the predicted spectra were the peak/trough locations in the higher frequency region. This allows us to make hypotheses regarding /s/.

1. The lower frequency region (below 2.5 kHz for subject CS) does not carry the cues for the fricative's identification because the mismatch in the pole/zero frequencies between the predicted spectra is greater than 500 Hz, yet the sound was still correctly identified as /s/. Consequently, this implies that the region of the back cavity is not important or its geometry does not need to be accurately modelled.
2. The distinct peak at 2.8 kHz does not carry the cue to the fricative's identification, but affects the virtual pitch of the sound. The ear is sensitive to the location of this peak, and although it may be slightly shifted (perhaps within a range of 200 Hz), the sound will still be identified as an /s/ sound.
3. The rise in amplitude levels above 4.0 kHz (for subject CS) carries the important cue to the fricative's identity. The listener is not sensitive to the peak and trough locations in the region; thus, an approximation of the amplitude levels may also produce a "relatively good" sounding /s/ noise (with "relatively good" based on the rating given by the listeners).

The fact that each separate analysis made here has led us to the same conclusions for /s/ and /ʃ/ strengthens the derived conclusions for the sibilant fricatives.

7.7.2 Nonsibilant Fricatives

The discussion of the nonsibilant fricatives will be more involved, since some of the observations made from the listening experiments seem to contradict the findings of the qualitative and quantitative analyses of Chapter 6. ANOVA on the MvN and BvN results have shown that the listeners failed to identify the fricatives /f/ and /θ/ with their natural counterparts. In fact, for the Mermelstein-derived /f/, the listeners made their decisions according to the sound, but chose the *incorrect* category of sound (refer back to the mean values of Table 7.9). For the MvB test, the t-test confirmed that the listeners significantly preferred the Blum-derived sound of /(a)f/ as sounding most like the natural sound, while for /θ/, the area-derivation technique was insignificant.

For /f/, the results of the listening experiment are in fact in agreement with the results of the quantitative analysis. Recall the clear mismatch between the Mermelstein-derived /(a)f/ with the measured spectrum (refer to Figure 6.5, page 103): the degree of match was measured at 6.0%, the lowest of all the fricatives. The MvN test results suggest that the listeners also significantly perceived this: when presented with the natural /(a)f/ and the alternative natural fricative (i.e., either /θ, s, ʃ/), they *chose the alternative sound* because they significantly decided that the natural and Mermelstein-derived /(a)f/ do not sound at all the same. This is why the null hypothesis was rejected for /(a)f/; the listeners *did* make their decisions based on the identification of the sound. A separate ANOVA test on the mean scores of /θ, s, ʃ/ (of the MvN test) confirmed further that the differences in the means of the alternate fricatives were insignificant. But if the listeners did not identify the Mermelstein-derived /(a)f/ because of the poor match to the measured spectrum, why then was the listener not able to match the Blum-derived /(a)f/, which has the highest degree of match, to the natural /(a)f/?

A comparison of the mean values of the BvN results (refer to Table 7.10, page 138) for /(a)f/, clearly shows that the mean for natural /(a)f/ is the highest, even though it is not significantly higher than the means of the alternate sounds. This finding complements the observations made by (Harris 1958); that is, the frication period of the nonsibilant fricative does not contain enough information on the fricative's identity and that the listener will not be able to (correctly) identify the fricative from only the frication period, but will also need to hear the transitional period of the sound (Harris 1958). This characteristic is different to that observed for the sibilant fricatives, whose cues lie in both the frication and transitional periods of the sound (Whalen 1991).

Therefore, we have found that even template-matching of sounds does not allow the listener to identify correctly the nonsibilant noise. However, we are pleased to find that the listeners did significantly choose the Blum-derived /(a)f/ over the Mermelstein-derived /(a)f/ as sounding most similar to the natural /(a)f/ in the MvB test. This means that, though the Blum area-derivation technique does significantly produce a better-matched version of the sound, and although the frication period does not contain information (or enough information) on the identity of the sound, the matches in the spectra, possibly from the lower frequency region below 4.0 kHz, must have given the sound qualities which are similar to the natural sound.

When the side branches were added to the /(a)f/ and /(i)f/ models, recall that the effects on the predicted spectra were small. There were slight shifts of less than 50 Hz on some of the peaks and troughs above 4 kHz. These shifts are relatively small compared to the observed 200 Hz variation in the spectra of the individual segments (previously discussed in Section 3.6) . The quantitative analysis has shown that addition of the side cavities does not consistently increase the degree of match between the measured and predicted spectra. The results of the BvNB test show that these shifts are perceptually insignificant to the listener. This observation supports the earlier suggestion that the side branches do not play an important role in modelling fricative sounds.

Now turn attention to the nonsibilant /θ/. The results of the MvN and BvN tests both suggest that the listeners were not able to identify this fricative and the results of the

MvB test show that the area-derivation techniques for this fricative were insignificant to the listener (i.e., the same quality of sound was produced). These results are quite contradictory to the findings of the quantitative analysis: the relatively high degree of match to the measured spectrum suggests that they should be easily identified by the listeners and the higher degree of match for the Blum-derived spectrum suggests that the Blum-derived sound should be significantly more identifiable to the listener.

In the discussion of the previous chapter, mismatches in the nonsibilant spectra (particularly for the Blum-derived /(i)f, (u)f/ and the Mermelstein-derived nonsibilants in general) were attributed as potentially being caused by the limitations of ACTRA, lack of realism in radiation characteristic estimation and of the teeth-airway boundaries. However, the Blum-derived spectra of /(a)f/ and /θ/ matched relatively well the measured counterpart spectra. If the Blum-derived /(a)f/ was definitely identified to match the natural sound better than the Mermelstein-derived /(a)f/ in the MvB test, why then was the Blum-derived /θ/ not identified to match the natural sound better than the Mermelstein-derived /θ/? It is obvious that there are very good matches up to 6.0 kHz for the Blum-derived spectrum of /θ/ (refer to Figure 6.4, page 102).

Now suspicion arises as to why the listeners chose the Blum-derived /(a)f/ over the Mermelstein-derived /(a)f/ in the MvB test in the first place: perhaps it was not because the Blum-derived version matched the natural version more, but because the listeners significantly decided that the Mermelstein-derived version did not sound at all like the natural version. This line of reasoning is supported by the results of the MvN test: that that the listeners significantly chose the alternative choice for this sound. If this was the case, it would explain why the Blum-derived /θ/ was not selected by the listeners as sounding more like the natural sound, even though the predicted spectrum shows a good match to the measured spectrum. This inevitably brings us back to the same conclusion as before: that the cues to the fricative's identity are not found in the frication period alone. Additionally, the listener is not sensitive to precise matches of the peaks and troughs in the spectrum.

To further support the conclusions for the nonsibilant fricatives, compare the ratings given by the listeners for the fricatives /(a)f, (i)f, (u)f/ and /θ/ (shown in Table 7.6, page 129). All received similar scores, even the segments of sound that were naturally uttered. One would expect the Mermelstein-derived—completely flat—/(a)f/ to receive a relatively lower rating than the Blum-derived sound. But, as was seen in the table, this was not to be the case. The observation brings forth the following line of reasoning:

The listeners are sensitive to the existence of the peak/troughs of the lower frequency region (about 4.0 kHz). But the exact location of these peaks may not be important, since the listeners still failed to identify and pick the well-matched Blum-derived /θ/. The Blum-derived /(a)f/ was chosen probably because its Mermelstein counterpart sounded significantly different due to the lack of peaks/troughs in the spectrum. It may be speculated that it is the higher frequency region which is more important for identification: the energy in this region carries some cues to the fricative's identity, albeit that these cues (by themselves) do not provide enough information as to the fricative's identity. However, once the listener was informed of the category of the sound, even the "flat" approximation of the Mermelstein-derived spectra was an adequate representation of /f/. And because there were not enough cues to begin with (seeing as the presented sound came from the frication segment alone), the same rating was given by the listeners to all the nonsibilant fricatives, which is "poor". The scores of the RtS test also imply that in the higher frequency region of the spectra, the amplitude levels are perhaps more important to the listener than actual locations of peaks and troughs. Consequently, the Mermelstein-derived /(a)f/ although flat and devoid of any peaks or troughs—was still able to produce a synthetic /f/ sound that was comparable to the naturally uttered version.

7.7.3 Side Branches

The results of the BvNB test have indicated that the side branches did not significantly affect the decisions made by the listeners, with the exception of /ʃ/, in which case the listeners preferred the synthetic version derived without the inclusion of side branches (i.e., both sublingual cavity and the pyriform sinuses). In light of the conclusions that were reached with regard to the sibilant and nonsibilant fricative, these results are analysed more carefully to ensure that they are in agreement with the hypothesis.

For the sibilant fricatives, the addition of the pyriform sinuses introduced an additional peak at approximately 6 kHz. The listeners did not significantly perceive this change and decided it was no different from the version which was modelled without the side branches. From here it was speculated that the pyriform sinuses might not be important for /s/, which agrees with the hypothesis that the supra-laryngeal region of the vocal tract does not have to be modelled accurately. Additionally, the perceptually unimportant peak in the high frequency region supports the argument made about /ʃ/ and even for that of the other fricatives; that is, in the high frequency region of the spectra, the energy levels are more important to the fricative's identification than the precise location of the peaks and troughs.

For the nonsibilant fricatives, it was observed that the listeners did not perceive any change in the sound when the area of the sublingual cavity and pyriform sinuses were included in the modelling. This agrees with the earlier hypothesis that the listeners were not sensitive to the locations of the peak and trough frequencies of the nonsibilant spectra. In fact, even a flat spectral shape in the higher region is adequate to model the frication period of an /f/ sound.

Further discussion is needed of the importance of the sublingual cavity to the production of fricatives, because after all, it is located in the most anterior region of the vocal tract. For the sibilants, the sublingual cavity is possibly more important than for the nonsibilants, because as previously stated, the cue to the identity of the nonsibilant fricative does not lie in the frication period alone. But for the sibilants and particularly for /ʃ/, addition of the sublingual cavity significantly affected the location of the significant peak at 2.7 kHz and raised its amplitude in the Blum-derived spectrum of /ʃ/. The listeners significantly preferred the version without the sublingual cavity, whereas one would have expected that the addition of the sublingual cavity would result in an even closer match to the natural sound.

The hypothesised explanation for this is fairly simple. It is adequate to model the sublingual cavity as an increase in area as opposed to a side branch, which is why the Mermelstein derived version was preferred by the listeners as sounding more like the natural sound. This is most likely because the area of the sublingual cavity itself is small, and so it is very likely that the Blum technique will read in the same area *twice*; once when the grid plane is placed normal to the main longitudinal axis, and a part of it may cover some area of the side cavity, and a second time when the grid plane is placed on the longitudinal axis of the side branch and a part of it may extend into the main tract (depending on the angle). Refer to Figure 4.9 shown on page 68 for clarity.

Thus, it is concluded that side branches may not play a significant role in fricative modelling. The pyriform sinuses may be excluded from the model because, for the sibilants, the precise geometry of the back cavity region is of lesser importance, and for the nonsibilants, the listeners are not sensitive to the peak and trough frequencies of the spectra. The sublingual cavity is more important for /ʃ/ since it affects the position of the (only) significant peak in the spectrum. But it is adequate to model it simply as an increase in area, possibly because of its shape and/or size. For /s/, the sublingual cavity is probably of lesser importance: its formation is not consistent across speakers, unlike /ʃ/; certainly for subject CS it was nonexistent, suggesting that it is not necessary to model the sublingual cavity as a separate branch to produce an /s/ sound.

Finally, it must be emphasised that there were many assumptions made about the modelling of the side branches that must be taken into account, as they may have considerably affected the outcome of the results just analysed. For example, the pyriform sinuses were modelled as a single orifice instead of two cone-shaped branches of the tract. Baer et al. (1991), in an investigation pertaining to vowels, stated that even if the dimensions of the vocal tract were accurately modelled, "the lack of available knowledge about the mechanical properties of the pyriform membranes and the degree of mechano-acoustic coupling that exists between the larynx tube and adjacent sinuses makes it difficult theoretically to examine their acoustic effects with any conviction", which is probably why Mermelstein excluded side branches from his model.

7.8 Summary and Conclusions

In this chapter, synthetic fricative sounds were created by taking the combined product of the spectral envelope of the predicted spectra, derived in Chapter 6, with the power spectrum of a noise source and applying the overlap-add technique to the synthetic waveform to create a segment of fricative sound. Natural segments of sound were created from the utterances of the speech corpus using the same technique for consistency. The sound segments were then used in a listening experiment consisting of five separate tests known as the MvN, BvN, MvB, BvNB and RtS tests. The first four tests are known as an 2AFC test, which requires the listeners to match a given test sound to one of two other sounds. The final test was a rating scheme for each synthetic and natural fricative segment.

Statistical analysis was applied to the results of the MvN, BvN, MvB and BvNB tests to determine whether the differences in the scores obtained from the listeners were statistically significant and whether there is sufficient evidence to conclude that the listener's decisions were based on identification of the fricative sound, the area-derivation technique, the inclusion of side branches or at random. Two types of tests were used, the *t*-test and the analysis of variance (ANOVA).

The tests revealed that, at the 95% level of significance, the listeners significantly identified the category of the sibilant fricatives /s/ and /ʃ/ but failed in the case of the nonsibilant fricatives /f/ and /θ/. It was also discovered that the listeners significantly preferred the Mermelstein-derived /ʃ/ and the Blum-derived /(a)f/. For the side branches, it was found that the listeners significantly preferred the version of /ʃ/ without side branches, while for fricatives /s, (i)f, (a)f/ addition of the side cavities did not have any significance for the listeners. The RtS scores showed that each type of fricative received similar ratings, indicating that the listeners found the synthetic sounds comparable in quality to each other and to the natural versions.

Using the results of the listening tests, it is hypothesised that the listener was able to identify the category of the sibilant fricatives based on the frication sound, but, for the nonsibilant fricatives, the cues for their identification do not lie in the frication period. This is because the listeners significantly failed to identify the nonsibilant fricatives even though the predicted spectra showed high degrees of matches to the measured spectra, even higher than that of the sibilant fricatives. This finding is in agreement with the results of Harris (1958), who observed that the nonsibilant fricatives cannot be identified from only the frication period of the sound, unlike the sibilant fricatives. The results of the listening experiment, along with careful observations of the predicted and measured spectra, have allowed us to derive a hypothesis on the role of each region of the fricative spectra.

For the sibilant fricatives:

1. the cues which allow the listener to identify the fricative are contained in the high frequency region of the fricative;
2. in this region, the amplitude levels are more important; the listeners are not sensitive to peak and trough frequencies;
3. the significant peak in the vicinity of 2.7 kHz gives the sound a pitch quality which is significantly recognised by the listeners; and
4. the lower frequency region is insignificant to the identification of the fricative and thus accurate measurements of the back-cavity are unnecessary.

For the nonsibilant fricatives:

1. the low frequency region does not contribute to the identification of the fricative;
2. any matches in the peaks and troughs of the lower frequency region are not useful to the listener if the listener does not know the identity of the fricative; and
3. in the higher frequency region, the amplitude levels are more important; listeners are not sensitive to peak and trough frequencies in this region.

For the side branches:

1. the pyriform sinuses are not important because they are located in the posterior region of the vocal tract; and
2. the sublingual cavity is only important for /ʃ/ but it is better to model it as an abrupt increase in area.

The next chapter uses the results of this analysis along with the findings from the previous chapters to determine the best area derivation technique for fricative synthesis. The final conclusions provide answers to the main objectives specified in Chapter 1.

Chapter 8

Discussion and Conclusions

In this chapter, the results of the previous chapters are brought together and discussed to make the final conclusions. From Chapter 1, the main objectives to accomplish from this investigation were:

1. to determine the perceptually important features of the spectra;
2. to determine the importance of the side branches;
3. to determine the best area-derivation technique for fricative synthesis; and
4. to determine whether MRI images are suitable for a "dynamic" study of the fricative.

Before discussion of the overall outcomes and final conclusions, the main procedures of this thesis and the results are summarised.

8.1 Summary of Investigation and Procedures

It has always been a mystery to speech scientists how the brain is able to interpret the vast number of cues within the speech signal (including identifying fricative sounds). The lack of knowledge on this subject was the main motivation behind this investigation. To tackle the problem, an acoustic and articulatory approach was used. Since much emphasis has been placed in previous investigations on the need for accurate information on the geometry of fricative articulation, MR images were used to study the articulatory properties of the vocal tract during fricative production. The images are volume scans of high-resolution, and are amenable to 3-D processing. This allows derived measurements of the cross-sectional area directly from 3-D models of the tract without any kind of transformation. Because much emphasis has also been placed on vowel context and how formant transitions may assist in the identification of the fricative, the speech and image data were recorded/captured while the speaker sustained the fricative between two vowel contexts.

There were several techniques reported in the literature as to how the area functions could be derived. It was found that two in particular were suitable to implement on MR images: the Mermelstein technique and the Blum transform. These two techniques were significantly different from each other with regard to the derivation and use of the longitudinal axis of the vocal tract and the treatment of side branches. Because of this, it was decided to use both techniques to derive the area function and to find out which one worked best for modelling fricatives.

The first part of this experiment investigated the spectral energy in the acoustic spectra. The PSD for eight consecutive segments within the fricative steady state of a single [VFV] utterance were derived to obtain the averaged PSD. From here, the distinct features of each spectrum were noted, the peak and trough frequencies and the differences in amplitude levels across fricatives. The variation of the spectral features was also studied by overlaying the PSD of the eight consecutive segments on a single graph for comparison. In the final part of the analysis, a procedure to quantify the degree of vocalic influence on each fricative was defined and the results compared with findings from the literature.

In the second part of the experiment, the image corpus was analysed. The 3-D models of the head and tract were created from the MR image scans, and boundaries between the teeth and vocal tract airway were determined from a procedure utilising cubic-shaped moulds of the subject's dental impressions. The area derivation techniques were implemented on the models using 3D-Doctor, CorelDraw and some manual computation. The side branches of the Blum-derived area functions were graphed separately from the areas of the main tract, because the area function would be modelled in two ways: first without the side branches, so that they could be compared to the Mermelstein-derived spectra, and secondly, with the areas of side branches so that the effects on the spectra as a result of their addition could be studied. The area functions for each technique were compared to each other and the differences between them were noted. Specifically, for the nonsibilant /f/, the area functions of the fricative sustained in different vowel context were also compared to investigate the effects of coarticulation during production.

In the third part of this experiment, the predicted spectra from the area functions were derived. ACTRA was used to achieve this, after meticulous testing beforehand for accuracy using simple models (from which could be calculated the output for comparison with the output of ACTRA) and complex models from the literature, whose measured spectra had been published. Several transfer functions were calculated from a single area function, each with the intermediate pressure source located at a different position. The transfer functions were multiplied by a radiation characteristic term and superpositioned to derive the predicted spectra, which approximates to the fricative spectrum for a distributed source model. The predicted spectra were compared to each other and to the measured (averaged PSD) spectra derived earlier, and differences and similarities between them were noted. The quantification procedure that was previously used to quantify the effects of vowel context was applied to the predicted and measured spectra of each fricative to calculate the degree of match between the spectra.

The fourth part of this experiment derived synthetic fricative sounds by taking the product of the predicted spectra with the power spectrum of the noise source. To obtain a final synthetic sound that had a longer duration, the derived waveform was processed using the overlap-add technique. For consistency, the natural segments were also processed using the same technique. The natural and synthetic sound segments were used in a listening experiment, which consisted of five separate tests and was conducted on 18 listeners. Each test had a different objective and utilised different combinations of synthetic and natural sounds. The tests are referred to as the Mermelstein versus Natural (MvN) test, Blum versus Natural (BvN) test, Mermelstein versus Blum (MvB) test, Branches versus No Branches (BvNB) test and Rate the Sound (RtS) test. In the MvN and BvN tests, the listeners were asked to identify the category of the synthetic fricative. In the MvB and BvNB test, the listeners were asked to choose the version of the synthetic fricative that best matched the natural counterpart. In the RtS test, the listeners were asked to rate the sound on a scale of one (very good) to five (very poor). The *t*-test and ANOVA statistical tests were applied to the results of the listening experiment to determine whether the decisions made by the listeners were based on identification of the sound or influenced by the area derivation technique or purely by chance or sampling errors.

8.2 Findings of Investigation

This section briefly describes the findings of the investigation. For convenience and clarity, this section is divided into 5 subsections.

8.2.1 Spectral Energy in Fricative Spectra

Analysis of the acoustic spectra has shown that the acoustic spectra of the sibilant fricatives have distinct features, which can be used to distinguish them from each other and from the nonsibilants. They are characterised by well-defined peaks and a large dynamic range. For /ʃ/, there is a significant peak located at 2.7 kHz and high spectral energy above 2.5 kHz. For /s/, there is a sharp peak at 2.7 kHz and steady increase in spectral energy above 4.0 kHz. By studying the spectra of the individual segments, the exact location of the peak/trough frequencies were found to be dispersed by approximately ±200 Hz from the location of the peak/troughs of the averaged spectra. It was observed that there was less variation in the vicinity of the aforementioned sharp peak of /s/ and the significant peak of /ʃ/.

The nonsibilant fricatives on the other hand showed similarities in spectral shape and have low dynamic range, making it hard to distinguish them from each other. The dispersions of peak/trough frequencies in the individual spectra were consistent throughout the spectra, contributing to the flatness of the spectral amplitudes, particularly in the region above 3.0 kHz. Differences between the averaged spectra of the nonsibilant fricatives seem to occur above 7.5 kHz, but, with analysis of only the spectral energy up to 8.0 kHz, only a small region of this part of the spectra could be seen. The averaged spectrum of /θ/ is similar to /(u)f/ because the energy level above 7.5 kHz decreases slightly. For /(a)f/, energy levels in this region increased slightly, while for /(i)f/ the spectra continued to be relatively flat.

8.2.2 Influence of Vocalic Context

The effects of vocalic context on the acoustic spectra were quantified using a procedure utilising mean square error measurements and dynamic time warping. The degree of match between two spectra is inversely proportional to how much the fricative is affected by vowel context; thus the higher the match, the less it is affected by vowel context. Although the quantification of the degree of vocalic influence on the fricative has never been reported in this manner, and the order of the "most affected-to-least affected" fricative never explicitly given, the results of the analysis were consistent with knowledge from the literature:

1. /f/ was most affected by vowel context and matches between the spectra were particularly low for /i/-/u/ contexts;
2. /s/ was the second most affected by vowel context and matches in the spectra were particularly low for /u/ context;
3. /θ/ was third most affected and matches between the spectra were particularly low for /u/ context; and
4. /ʃ/ was least affected by vowel context with similar degrees of match across the different vowel contexts.

The nonsibilant /f/ is known to be highly influenced by vowel context, because the tongue does not need to be in any particular position to form the constriction. Thus, its position is most highly influenced by the vocalic environment. The sibilant /s/ comes second; the effects of lip rounding have been known to extend into the acoustic spectrum of the sibilant /s/, probably as a result of a second source mechanism due to the whistle-like geometry of the lips in /u/ context (Shadle and Scully 1995). The nonsibilant /θ/ is sensitive to small shifts in constriction shape because of the very short front cavity, but less sensitive than /f/ because the tongue must be placed between the upper and lower incisors to form the constriction. And /ʃ/ is least likely to be affected by vowel context because the tongue blade must be elevated and rested against the palatal vault to form the long channel for the constriction, and because the teeth act as an obstacle in the path of the jet stream, producing high amplitude spectra with few obvious peaks/troughs (Shadle 1985).

Subsequently, the area functions of /f/ were also studied in vowel contexts /a, i, u/ to determine if coarticulatory effects were reflected in the area function. The results showed that they were; with the exception of the constriction region, the areas of tract were similar to the geometries of the vowels in which the fricative were uttered. These findings are consistent with those obtained by Engwall and Badin (2000), who measured coarticulatory effects in fricative articulation, and Shadle et al. (1996), who observed that the effects of vowel transitions were preserved in the sustained fricative segment. The question remains as to whether these coarticulatory effects are able to provide the cues to the fricative's identity. This issue will be further discussed in the following sections.

8.2.3 Differences in Mermelstein and Blum-Derived Area Functions and Spectra

In general, the Mermelstein- and Blum-derived area functions were different from each other. The major difference lies in areas of the laryngeal region where in some cases the Mermelstein-derived area functions show very small measurements while the Blum area-derivation technique was able to derive more consistent back area measurements across the different fricative geometries. Examples of this can be seen by comparing the area functions of /ʃ, (a)f, (u)f, θ/. The total lengths of the area functions were also different; the total length of the Blum-derived area functions was longer, because the grid planes were placed at intervals along an irregular longitudinal axis, while the Mermelstein area-derivation technique placed grid planes that were parallel to each other. One other difference between the area functions was the representation of the side branches. The Mermelstein-derived areas did not incorporate side branches, unless the area was directly connected to the main airway on the measured grid plane. This means that the side branches were modelled as an increase in area. For the Blum-derived area functions, the side branches were presented separately from the main tract, because they were derived using a separate longitudinal axis. Consequently, for the modelling, the side branches of the Blum-derived area functions were realistically modelled as a side branch.

The pyriform sinuses were derived for all the fricatives with the exception of /(u)f/. In this case, they were located on either side of the vocal tract and could not be observed from a midsagittal view. The derived areas of the pyriform sinuses varied significantly. This is quite common across different phonemes and across speakers and has been noted in other investigations utilising MR images (Baer et al. 1991; Narayanan and Alwan 2000). The sublingual cavity was derived for all the fricatives with the exception of /s/ and /θ/. For /θ/, this may be expected because of the unique tongue location but for /s/, it implies that subject CS had placed her tongue against the lower incisors to form the constriction.

The Mermelstein and Blum-derived spectra of /ʃ/ show relatively small differences given the significant difference in the area measurements of the laryngeal region. The significant peak at 2.7 kHz was correctly modelled in the Mermelstein-derived spectrum, the same peak was lower in frequency and amplitude in the Blum-derived spectrum. Small perturbations occurred in the lower frequency region of the Blum-derived spectra, possibly attributable to the incomplete cancellation of the resonances of the back cavity. In the region above 3.0 kHz, both predicted spectra show a smooth spectral shape with no obvious peaks or troughs.

The predicted spectra of /s/ show peaks at various frequencies and bandwidths in the region above 4.0 kHz, which differ between the two spectra. The sharp peak at 2.8 kHz was correctly modelled in the Mermelstein-derived spectrum; however, in the Blum-derived spectrum it was lowered by about 200 Hz. Both predicted spectra show obvious mismatch from the measured spectrum in terms of the frequencies of the first pole and zero pair. The consistency in the measurements of the area functions and consequently the derived spectra suggests that the mismatch of this first pole/zero pair may perhaps be caused by differences in the subject's articulatory positions during imaging and recording sessions.

The Blum-derived spectrum of /θ/ shows well-matched peaks with the measured spectrum of /θ/. Between 2.5 kHz and 6.0 kHz the Blum-derived spectrum is almost like the spectral envelope of the measured spectrum. Before and after this frequency range, there were differences in peak/trough frequencies, but the mismatches appeared small. On the other hand, the Mermelstein-derived spectrum did not show good matches to the measured spectrum. The first pole and zero pair was completely cancelled and the second trough was lowered. The only match in the high frequency region was the peak at 4.3 kHz.

The three vowel contexts for the labiodental /f/ show major differences even between spectra derived using the same technique. It is possible that the effects of coarticulation could have been attenuated by the inaccuracies of the area functions or, when the pressure sources are located at the inner edges of the lips, the assumption of one-dimensional wave propagation breaks down. Nevertheless, the Blum-derived area function managed to produce a good prediction for /(a)f/. Except for some slight lowering in frequency, the shape of the peaks match well to the measured spectrum, particularly in the region below 4.0 kHz. The Mermelstein-derived spectrum on the other hand, was flat, as a result of complete cancellation of the pole and zero frequencies. In the case of /(u)f/, the Blum-derived spectrum again gave better matches in the lower frequency region and for /(i)f/ both the predicted spectra looked similar. This was quite expected, since the area functions of both techniques gave similar area measurements across the vocal tract length. Similar to /s/, it is postulated that the mismatch in the first zero in the spectra may be caused by the different positioning of the articulators during the imaging and recording sessions.

8.2.4 Quantified Differences Between Measured and Predicted Spectra

The quantification process was applied to measure the degree of match, D_m, between the predicted and measured spectra. It is assumed that if the calculated D_m value of two spectra were relatively high, then the area-derivation technique would have succeeded in deriving relatively good area measurements. The averaged D_m values of the Mermelstein-derived spectra, the Blum-derived spectra without side branches and the Blum-derived spectra with side branches showed that the Blum-derived spectra without side branches best matched the measured counterpart. The Blum-derived spectra modelled with side branches came in second, although the difference in the averaged D_m values did not seem significant, and the Mermelstein-derived spectra came in last with a significantly lower averaged D_m value. However, the Mermelstein technique did produce the highest degree of matches for the sibilant fricatives.

In terms of type of fricative, the sibilant fricatives had the best predictions, while the nonsibilant fricatives were less well-matched. The fact that the nonsibilant fricatives were harder to model has been noted before by Narayanan and Alwan (2000). The very short front cavities make them sensitive to slight changes in the constriction shape. In terms of side branches, there was no clear indication as to whether their addition produced better-matched predictions, because the degree of match improved in some cases and worsened in others. For specific cases, the best matched spectrum was that of the Blum-derived /(a)f/ and the worst matched spectrum was that of the Mermelstein-derived /(a)f/. This is consistent with the features of the spectra and particularly so in the case of the Mermelstein-derived /(a)f/, which was completely flat with no peaks or troughs.

8.2.5 Significant Results of Listening Experiment

There were five tests in the listening experiment, each part having a different objective. The first test (MvN) required the listener to identify the category of the Mermelstein-derived fricative sounds. The second test (BvN) required the listener to identify the category of the Blum-derived fricative sound. The third test required the listener to choose the preferred area derivation technique. The fourth test required the listener to pick the model derived with or without side branches. For these tests, the listeners were not aware that the sounds were synthetic and natural segments of speech. The fifth test required the listener to give a goodness measure for the fricative segment on a scale of one (very poor) to five (very good). The first four tests mentioned above are known as "two-alternative forced-choice" tests and the order of the tests was randomised across subjects.

1. In the MvN test, the listeners successfully identified the correct categories of the fricatives /s/ and /ʃ/. The listeners mistakenly identified /(a)f/ as either /θ, s, ʃ/. The fricative /θ/ failed to be identified.

2. In the BvN test, the listeners successfully identified the category of the fricative /ʃ/. The listeners mistakenly identified /s/ as /ʃ/ almost half of the time. The nonsibilant fricatives failed to be identified.

3. In the MvB test, the listeners identified the Mermelstein-derived /ʃ/ and the Blum-derived /(a)f/ as sounding more similar to the natural sound. There was no preference over the Mermelstein- or Blum-derived sounds for /θ/ and /s/.

4. In the BvNB test, the listeners decided that the version for /ʃ/ modelled without side branches sounded more similar to the natural sound. There was no preference for /s, (a)f, (i)f/.

5. In the RtS test, each fricative category received similar scores regardless of the type of sound (i.e., synthetic or natural). The averaged and rounded scores for each fricative are: 4 for /ʃ/, 3 for /s/, and 2 for both /θ/ and /f/.

8.3 Discussion and Final Conclusions

We are aware of the fact that the area-derivation technique of Mermelstein was applied without any modifications whatsoever. Other investigations utilising this technique have angled the plane posterior to the vocal tract bend to match the vertical axis of the laryngeal region (e.g., Badin et al. 1998). Instead, the angle of the bend was kept at exactly 90°. This was done on purpose because as was previously mentioned, two distinctively different area-derivation techniques will produce variations in the spectra, which may or may not prove to be perceptually important to the listener.

The area functions, the predicted spectra and the results from the listening experiment have allowed speculation on the perceptually important features of the acoustic spectra of fricatives. Below is a brief description of the conclusions.

The sibilant /s/: The low frequency region, and specifically the sharp peak at 2.8 kHz, does not contribute to the identification of the sound. Instead it gives the sound the pitch quality which was significantly perceived by the listeners. If the peaks in this region do not match the peaks in the measured spectrum (e.g., such as the mismatch in the first pole-zero pair at 800 Hz), the sound can still be identified as /s/. The amplitude levels in the frequency region above 3.0 kHz are the spectral cues to the fricative's identity. It is characterised as a steady increase in amplitude, reaching the maximum level at 7.8 kHz (for the 8 kHz frequency range of interest). The predicted spectrum does show differences in peak frequencies in this region. The listeners are not sensitive to their frequencies; the addition of the pyriform sinuses introduced additional peaks in the region which were not perceived by the listeners. Thus, it is highly possible that the cues to the fricative's identity depend entirely on the relative energy levels between the lower and higher frequency region, with the amplitude increasing steadily.

The sibilant /ʃ/: The low frequency region below 2.0 kHz is not important to the listeners; the major differences in the area functions had only small acoustic effects in this region. The significant peak at 2.8 kHz is the first front cavity resonance and gives the sound the pitch quality that is perceived by listeners. The higher frequencies carry the cues to the fricative's identity. There are no obvious peaks or troughs in this region (consistent with the measured spectrum); therefore the cues to the fricative's identity is in the energy level alone. It is characterised by constantly high energy levels above 3.0 kHz.

The nonsibilant /θ/: The identity of the fricative does not lie in the frication period. The listeners cannot identify the fricative based on the sound. Any matches in the lower frequency region are not important to the listener. A spectral envelope of the high frequency region is adequate to produce a version of the sound which is comparable to the naturally uttered sound.

The nonsibilant /f/: The identity of the fricative does not lie in the frication period. The listeners cannot identify the fricative based on the sound. Any matches in the lower frequency region are not important to the listener. A spectral envelope of the high frequency region is adequate to produce a version of the sound which is similar to the naturally uttered sound.

It was speculated that accurate measurements of the back cavity region are not important for the strident fricatives. For /ʃ/, the Mermelstein-derived area produced low and irregular measurements in the laryngeal region, yet this big difference had only a small effect on the predicted spectrum, while for /s/, the mismatch in the location of the first pole-zero pair did not deter the listeners from identifying the fricative. The cues to the fricative's identity are suspected to be mostly dependent on the higher frequency region. The geometric region of interest for fricative modelling is therefore only from the lips to the region of constriction. But accurate measurements in this region may not be very important for the sibilants either. A configuration that will produce good approximation of the amplitude levels in the higher frequency region (above 2.5–3.0 kHz) is enough to derive a "good" identifiable version of the /ʃ/ sound and a "relatively good" identifiable version of an /s/ sound (based on the ratings given by the listeners). This is because the presence of an obstacle (such as the teeth) in the path of the jet stream results in a high-amplitude spectrum with few obvious poles and zeros (Shadle 1985). It was observed that the exact location of these poles and zeros are not perceptually important to the listener. This conclusion is based on two observations: the addition of the pyriform sinuses into the /s/ model introduced a new peak at 6.0 kHz in the Blum-derived spectrum but this addition was insignificant to the listener; and the Mermelstein- and Blum-derived /ʃ/ spectra were relatively flat above 3.0 kHz and obviously lacked some of the features of the measured spectrum, such as the broad peak at 4.0 kHz. Yet both synthetic sounds were correctly identified as /ʃ/ and given a rating that was even higher than the natural sound.

It is noted that the peak in the vicinity of 2.8 kHz of the sibilant fricatives is important. It gives the sound a pitch quality, which is significantly perceived by the listeners, allowing them to match the sound to the natural version. It is speculated that the Blum-derived /s/ was mistakenly identified as /ʃ/ because, coincidentally, the sharp peak at 2.7 kHz was located at the same frequency as the (only) significant peak of measured /ʃ/. This reasoning comes from the fact that some listeners admitted to making their choices based on pitch quality alone, since they did not know what the sounds were supposed to represent. Recall too the results of the BvNB test: addition of the sublingual cavity and the pyriform sinuses sharply increased the amplitude of the peak at 2.8 kHz. The listeners perceived this change and significantly preferred the version of sound without the side branches, supporting the hypothesis on the significance of the peak in this vicinity.

For the nonstrident fricatives, it is evident that there were not enough cues in the sound to help the listener identify the type of fricative. This is supported by the fact that the Blum-derived /(a)f/ and /θ/ sound were unidentified by the listeners, even though the degrees of match between the predicted and measured spectra were even higher than the sibilant fricatives. In fact, the Blum-derived spectrum of /θ/ was almost like a spectral envelope of the measured spectrum. But the listeners failure to identify nonsibilant fricatives is in fact consistent with the findings of Harris (1958), who observed that naturally-uttered nonsibilant fricatives could not be identified if the vocalic segment of the sound was uttered in the context of another fricative (indicating the importance of the transitional period between the fricative and vowel). It was noted, however, that the Blum-derived /(a)f/ was preferred by the listeners in the MvN test. Initially, it was concluded that the Blum-derived /(a)f/ was identified by the listeners because of the good match in the lower frequency region, which implied that, similar to the sibilant fricatives, it must have given the sound a pitch quality that matched the natural sound and was significantly perceived by the listeners. However, this does not explain why the Blum-derived synthetic /θ/ sound was not preferred in the MvB test or why the fricative scores were similar even for the low-matched Mermelstein-derived spectrum. Therefore, it was decided that there might be another explanation as to why the Blum technique was preferred for /(a)f/ and it was speculated that the Blum-derived /(a)f/ was chosen simply because the Mermelstein-derived /(a)f/ sounded so significantly different from the natural /(a)f/ sound, because of the lack of peaks/troughs in the spectrum. This new line of reasoning supports the results for /(a)f/ of the MvN test, where the listeners significantly chose the alternative fricative instead of natural /(a)f/. At the same time, it explains why the area derivation techniques for /θ/ were insignificant to the listener for the MvB test: the listeners are less sensitive to the peaks and trough frequencies of the nonsibilant sounds and cannot tell the sounds apart based on their frequencies. The consequence of this reasoning is that it also implies that the matches in the spectrum of the Blum-derived /(a)f/ and /θ/ are not important to the listener. Since the results mostly support the latter line of reasoning, it is in fact suggested that matches in lower frequencies of the nonsibilant fricative may be insignificant. Regardless of whether they match or not, they do not help the listener to identify the fricative nor match it to the template sound. Therefore, it is also concluded that a good approximation of the high frequency region is adequate to model the fricative, which is why the flat Mermelstein-derived /(a)f/ was also perceived as a "poor" version of /f/ along with other /f/ sounds, and in fact in some cases, received a slightly higher score than the natural and synthetic Blum-derived sound.

If only the amplitude levels in the high frequency region are in fact the cues to the fricative's identification, then the results here support the observation from the investigation by Stevens (1985), who, by changing the ratio of the noise level with regard to the amplitude level in the higher frequency region of the neighbouring vowel, created identifiable synthetic versions of the /s, ʃ, θ/ sounds. This means that the fricative must be uttered in a vocalic context, so that the transitions may provide the basis with which to compare the degree of noise. This line of reasoning also then supports the findings of Whalen (1991), who observed that the transition regions also provide cues for the identification of the sibilant fricatives.

So far it does seem as if the Mermelstein technique—although relatively easier to implement—has allowed it to derive the best predictions of the sibilants, while the complexity of the Blum technique, which allowed good modelling of the nonsibilants has gone to waste! To be fair, both area-derivation techniques have their benefits and disadvantages. The Blum technique was attractive because the side branches could be incorporated into the model as separate areas, but the decrease in the degree of match after addition of the side branches for /ʃ/*, /ʃ/**, /(a)f/* and /(a)f/**, the perceptually insignificant role of side branches for the BvNB test, and the similar ratings given to sounds with and without side branches in the RtS test all suggest that the side branches are not a worthy addition to the model. Additionally, the fact that Blum was able to derive good matches in the lower frequency region of the nonsibilant fricatives could also have been achieved by the Mermelstein technique, if the vertical section of the longitudinal axis of the vocal tract were slightly angled so that it is in parallel to the laryngeal axis.

Thus the conclusion is this: the Mermelstein technique is best, because it is relatively simpler to apply and because there is no need to incorporate the side branches in the model. As suggested by the results, highly accurate measurements of the vocal tract geometry are unnecessary to produce a good version of the fricative sound, as the model itself is restricted and highly influenced by other aspects not directly related to the area function, such as the modelling of turbulent noise on the lip surface. Perhaps it is more accurate to suggest that high resolution MR images may not be necessary at all, and that a concatenation of tubes which replicate the main features of the oral region, including constriction, is adequate to model the synthetic sound. By changing slightly the areas of the front cavity (as was seen in the case of /ʃ/ when the sublingual cavity was added into the model), the significant peak of the sibilants can be moved about to a frequency that will give the sound a natural quality.

For the nonsibilant fricatives, it is perhaps more worthy to investigate the transitions of the sound rather than the sound itself. Although the effects of coarticulation were observed in the area functions and consequently reflected in the predicted spectra, they did not help the listener to identify the fricative. It is possible that the effects of coarticulation in the spectra were enhanced (but incorrectly) in the area function. Engwall (1992) had observed that articulation of the vowels was affected when they are sustained, causing them to differ slightly from normal conditions, particularly for /i, u/. This may have affected the predicted spectra of /(i)f/ and /(u)f/, since we did observe the effects of coarticulation in the area functions. This may explain why /(a)f/ context produced the higher degree of match to the measured spectrum than /(i)f/ and /(u)f/.

Perhaps the RtS scores given by the listeners say it all, where the natural derived sound in some cases received an even lower rating than the synthetic sound, and there were no significant differences in the scores of /f/ uttered in different vowel contexts. These results indicate that the effects of coarticulation may not be helpful for fricative identification, if only the steady state section of the noise is studied. Thus, static MR images of fricative steady state may not be the best approach to tackle the problem if intending to study the dynamic aspects of the fricative.

Before ending this section, the objectives and the findings of this investigation are summarised:

1. *To determine the perceptually important features of the spectra:*
   It was observed that the relative amplitude levels between the lower and upper frequency region provided the cue to the identity of the fricative. Specifically for the sibilant fricatives /s, ʃ/, the peak in the vicinity of 2.7 kHz gave the sound a pitch quality that was perceived by the sensitive ear of the listener and gave the sound a natural human-like quality. For the nonsibilant fricatives, the peaks and troughs of the lower frequency region did not contribute to the identification of the fricative or the pitch quality of the sound.

2. *To determine the importance of the side branches:*
   The pyriform sinuses are not important to the synthetic fricative model as they were perceptually insignificant to the listeners. The sublingual cavity was also observed to play a insignificant role, with the exception of /ʃ/. Slight changes in the area may contribute to changes in the amplitude and frequency of the significant peak at 2.7 kHz which, as mentioned before, affects the pitch quality. However, since this effect does not contribute to the fricative's identity cue, it is adequate to model the sublingual cavity as an increase in area, as opposed to a separate side branch.

3. *To determine the best area-derivation technique:*
   The Mermelstein technique is recommended, because it is easier to implement and because it is unnecessary to model the side branches as separate ducts. However, approximations of the vocal tract geometry from the lips to the constriction may also be adequate to produce the synthetic sound.

4. *To determine the usefulness of MRI images for a "dynamic" study of the fricative:*
   MR images of sustained nonsibilant fricatives in different vowel contexts may not be the best approach to tackling the problem from a dynamic angle. The effects of coarticulation, although preserved in the images, were not sufficient to provide the cue to the fricative's identity for the listener. It is suggested that a study of the articulation at transitional periods of the fricative-vowel will perhaps be more insightful. However, this may not be feasible at the moment, with the current MRI technology.

8.4 Suggestions for Future Work

We are aware of the fact that the images and recordings were made by an American speaker, yet the listeners of the experiment were British speakers. There is also the possibility that the average UK speaker may not place the tongue between the upper and lower incisors to form the nonsibilant /θ/ but almost at the edge of the alveolar ridge behind the upper incisors which can, for some speakers and to some listeners, make /θ/ sound more like a *soft* version of /d/. Perhaps a look at the articulation and acoustic spectra of /d/ in several vowel contexts may be one approach to determining the cue which differentiates the /θ/ sound from the /f/ sound.

This investigation has overlooked the importance of noise source models. The most important step to take next is perhaps to replicate this experiment and use a more realistic source model to check whether the conclusions remain the same. Additionally, the results of the listening experiment have given rise to a whole new set of questions that demand further investigation. It might be worthwhile to test whether the sibilant fricative can be identified solely from the cues of the higher frequency region. For instance if natural sibilant sounds were high-pass filtered at 3.0 kHz, would the listener still be able to identify the fricative? One suggestion for a future listening test is a synthesis of fricative sounds which only have spectral energy within a certain frequency range (e.g., 2–5 kHz and 5–7 kHz) for listeners to identify, or rate, as was done in the RtS test.

In the case of the side branches, the many variables and assumptions that were made in the effort to model them may have affected the outcome of the experiment. Although the region posterior to the constriction may not be important—which lessens the importance of the pyriform sinuses—the sublingual cavity may perhaps play a much more important role than speculated because it is located in the anterior region of the tract, very near to the constriction region. Therefore, it is a good idea to look into this further, again using more realistic noise source models.

The spectral comparison procedure has shown that it is a simple and reliable tool to measure the differences between two spectra. Perhaps the computed degree of match can be used as a parameter to distinguish the front fricatives /f/ and /θ/ from each other, since we know that the variations of /f/ in different vowel contexts are significantly larger than for /θ/.

Finally, the results of this investigation have strongly suggested that cues on identification of the nonsibilant fricatives are not dependent on the frication period even if coarticulatory effects are seen. At this moment in time, an investigation utilising volumetric MR images of the subject in a [FV] transition may not be possible, but advancements in technology may make it possible to capture 3-D images of subjects uttering complete [VFV] tokens. It is suggested that a look into the area functions during transitions from fricative to vowel may be the best approach to better understanding the cues that make the nonsibilant fricatives identifiable to listeners.

Appendix A

ACTRA: Preliminary Tests

A number of tests were carried out to compare the TFs calculated by ACTRA with standard models from the literature. These preliminary tests include calculation of the TF of each ACTRA element in isolation, vowel configurations and a uniform tube model with rigid and, later, reflecting walls and an intermediate pressure source. In these tests, the resonances of the TFs were compared with the resonances that were calculated using 1-D wave equations. Additionally, ACTRA was also tested with the mechanical flow duct model from Shadle (1985), which includes an obstacle as well as a pressure source. Finally, the TF of the area function of /s/ from an investigation by Narayanan and Alwan (2000) is predicted. For these tests, the TFs were compared to the measured spectra of Shadle (1985) and Narayanan and Alwan (2000), respectively.

A.1 Introduction

Throughout this appendix, reference is often made to the calculated percent error between the peaks/troughs of the TF calculated by ACTRA with the ones calculated using 1-D equations or observed in the measured spectrum. This error, E, is determined by computing the difference between the main poles and/or zeros of the ACTRA TF, denoted F_A, with the corresponding poles and/or zeros of the computed/measured spectrum, denoted F_M. The equation is given by:

$$E = \frac{F_M - F_A}{F_M} \times 100. \tag{A.1}$$

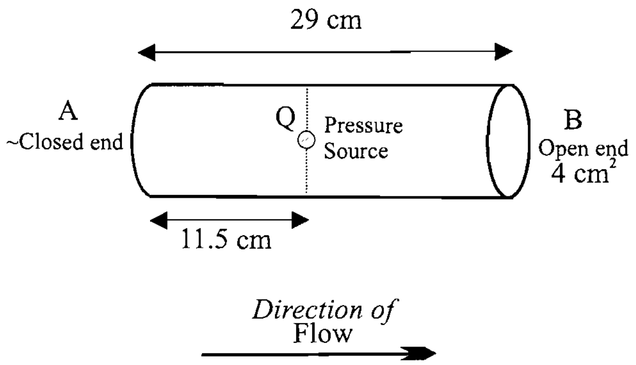

FIGURE A.1: Uniform tube configuration with intermediate pressure source.

Additionally, the 1-D plane wave equations for a uniform tube configuration are given by:

$$\text{Only open at one end: } F_n = \frac{c\,(2n+1)}{4l} \quad \text{(A.2)}$$

$$\text{Closed/open at both ends: } F_n = \frac{cn}{2l} \quad \text{(A.3)}$$

where c is the speed of sound, typically 353 m/s, l is the length of the tube and n is the formant number. The first few tests described here will compare the ACTRA-predicted formants with the formants computed using the above equations.

A.1.1 Uniform Tube Models

Each ACTRA element was first tested in isolation to check if the formants are consistent to the ones measured using the 1-D wave equations. This preliminary test was similar to that carried out by Jackson (2000) and will not be described in depth here. The averaged calculated error was observed to be less than 4.0%. Subsequently, two-tube configurations representing the vowels /a, i, æ/ were adapted from Flanagan (1972) and were modelled by ACTRA. The error between the ACTRA and calculated formants was again less than 4.0%. For the simulations mentioned here, the parameter list was altered to model rigid, non-reflecting walls with a speed of sound of 353 m/s, ratio of specific heats of 1.4, characteristic impedance of 402 kg/m^{2}s and a temperature of 273 K which assumes that the medium has 0% humidity.

Next, ACTRA was used to calculate the TFs for a uniform tube model with an intermediate pressure source. The configuration of the model is shown in Figure A.1. The tube measures 29 cm with a circular cross-sectional area of 4.0 cm^2. A pressure source is located in the tube at a distance of 17.5 cm from the open end.

Figure A.2 shows the TFs that were calculated by ACTRA for the given tube model for (a) the whole tube and, (b) from the closed end, A, to the pressure source, Q. The

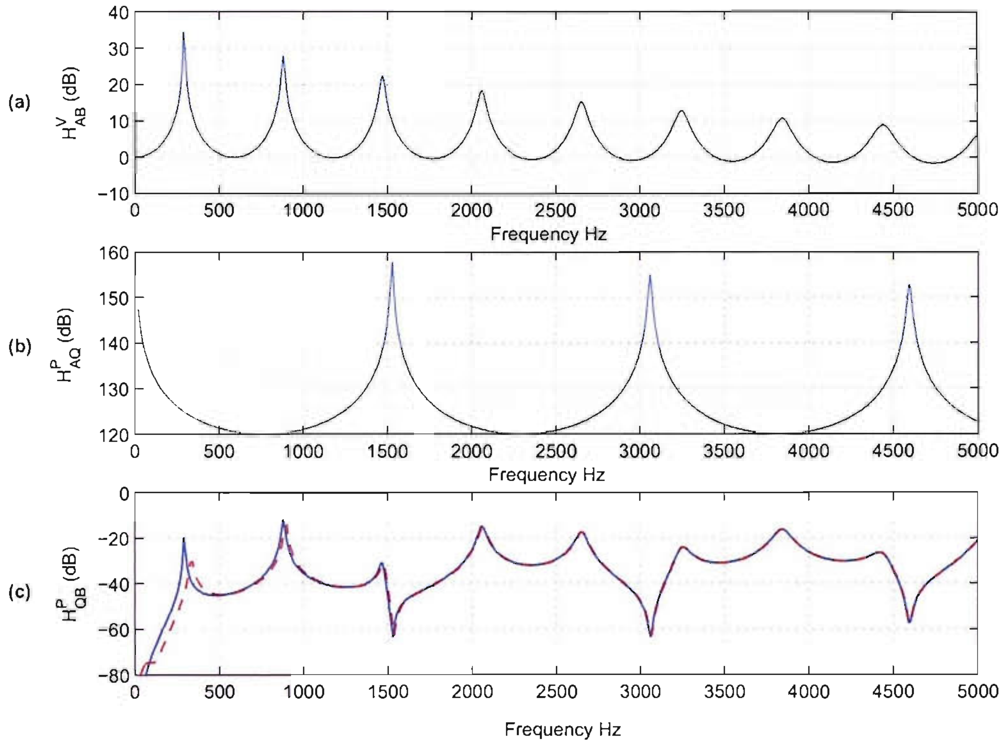

FIGURE A.2: Calculated TF for the uniform tube model with an intermediate pressure source with rigid walls. (a) H^V_{AB} of whole tube. (b) H^P_{QA} from closed end, A, to pressure source, Q. (c) H^P_{QL} downstream of pressure source. Red-dashed lines are H^P_{QL} for same tube model, this time with reflecting walls.

Poles of whole tube	F_0 (Hz)	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
Calculated	304	912	1520	2128
Predicted (ACTRA)	290	880	1470	2060
% Error	4.6	3.5	3.2	3.2
Zeros of rear cavity of tube	F_0 (Hz)	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
Calculated	0	1534	3069	4602
Predicted (ACTRA)	0	1530	3060	4590
% Error	0	0.2	0.2	0.2

TABLE A.1: Calculated and predicted formants for uniform tube model with intermediate pressure source.

reciprocal of (b) was computed and the combined product of the two TFs is shown in (c). The poles/zeros of the ACTRA TFs and the ones computed from the 1-D wave equations are given in Table A.1. The frequencies of the predicted formants of the ACTRA TFs were quite close to the calculated values: the calculated error was low, the maximum was 4.6%.

To see the effects of reflecting walls, the wall properties were adapted to the values of Table 5.2. The red-dashed lines in Figure A.2(c) show the calculated TF of the

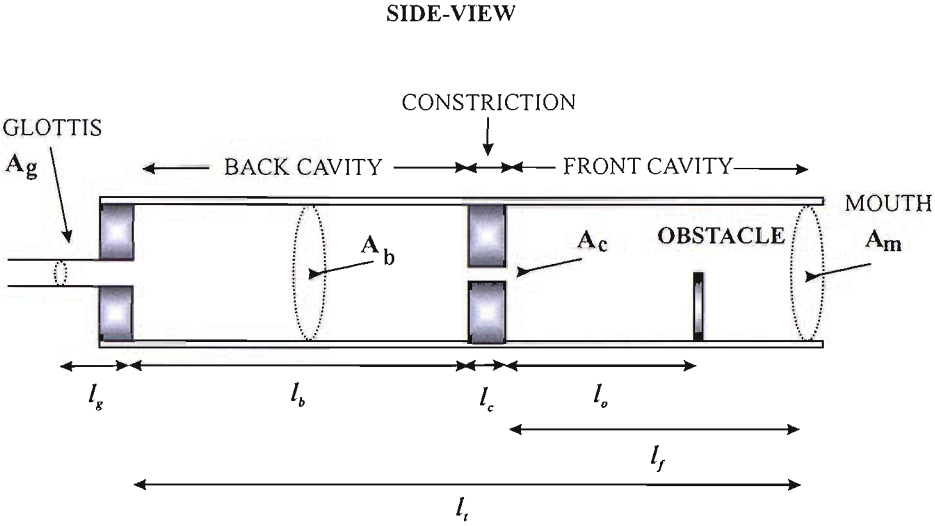

FIGURE A.3: Diagram of physical flow-duct model after Shadle (1985). Area and length are denoted by A and l respectively. See text for measurements.

same model with reflecting walls. Comparisons of the spectra in (c) show that the wall properties mainly affect the lower frequencies; there is a rise in the frequency of the first peak and second peak, more so for the first peak from 290 Hz to approximately 340 Hz. There is also a slight lowering of amplitude at the lower frequencies. This is consistent with the observation made by Davies et al. (1993). Adapting vocal tract wall properties raises the first formant frequency and increases its bandwidth substantially, but the frequencies of interest are greater than the wall frequency. Therefore the effects of wall vibration are seen to diminish quickly as the frequency increases.

A.1.2 Flow Duct Model with Pressure Source and Obstacle

ACTRA is next used to calculate the TFs of the flow duct model from Shadle (1985) which includes a pressure source and an obstacle downstream of the constriction. The original mechanical setup from Shadle (1985) is shown in Figure A.3.

The tube has a circular cross-section along its length, except at the location of the obstacle, which covers the bottom half of the tube. The obstacle was placed 5.07 cm downstream from the constriction (denoted l_o in figure). The length of the constriction, l_c, is 1.0 cm, with a cross-sectional area, A_c, of 0.079 cm^2. The length of the modelled glottis, l_g, is 2.0 cm, with a cross-sectional area, A_g, of 0.97 cm^2. The length of the back cavity, l_b, is 12.8 cm, with a cross-sectional area, A_b, of 5.067 cm^2.

The constituent properties adapted for this simulation followed the values used in Jackson (2000). The tube has rigid walls, with a volume flow-rate of 160 lit/s, speed of sound of 344.8 m/s, ratio of specific heats of 1.4, characteristic impedance of 402 kg/m^{2}s

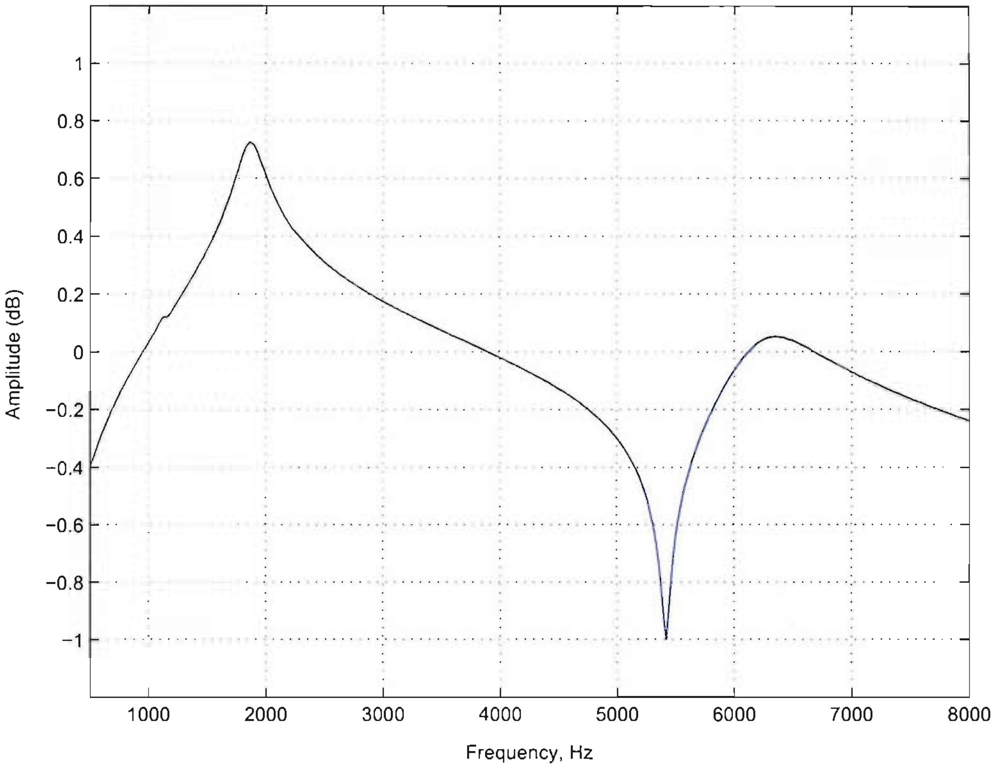

FIGURE A.4: Predicted spectrum for the flow-duct mechanical model from Shadle (1985). The spectrum has been normalised. See Jackson (2000), p. 46 for comparison.

and temperature of 293 K. The calculated TF was multiplied with the windowing function of the source model taken from Shadle (1985), given by ae^{bf}, where a and b are numerical constants with values of 80 and -0.0007 respectively for low flow-rates. The predicted spectrum is shown in Figure A.4.

The ACTRA derived spectrum has a shape similar to the measured response of Shadle (1985). The main features of the spectrum are the distinct peak at approximately 1.8 kHz, a distinct zero at approximately 5.4 kHz, followed by a broad peak at approximately 6.3 kHz. In Shadle's measured spectrum, the first distinct peak is located at approximately 1.8 kHz, the distinct zero at approximately 5.9 kHz and the broad peak at approximately 6.5 kHz. The averaged calculated error between these frequencies is approximately 4.0%, mainly because of the differences in the frequencies of the first zero and second broad peak. Jackson (2000) has noted discrepancies in ACTRA TFs in the higher frequencies (~6 kHz) which he attributes to poor estimation of the higher modes from multiplicative errors and other inaccuracies. However, he notes that the TFs calculated by ACTRA are still within the error bound of the classic electrical analogue predictions.

The models that were used to test ACTRA so far are relatively simple, because they consist of a maximum of three elements. The next section, ACTRA was used to calculate

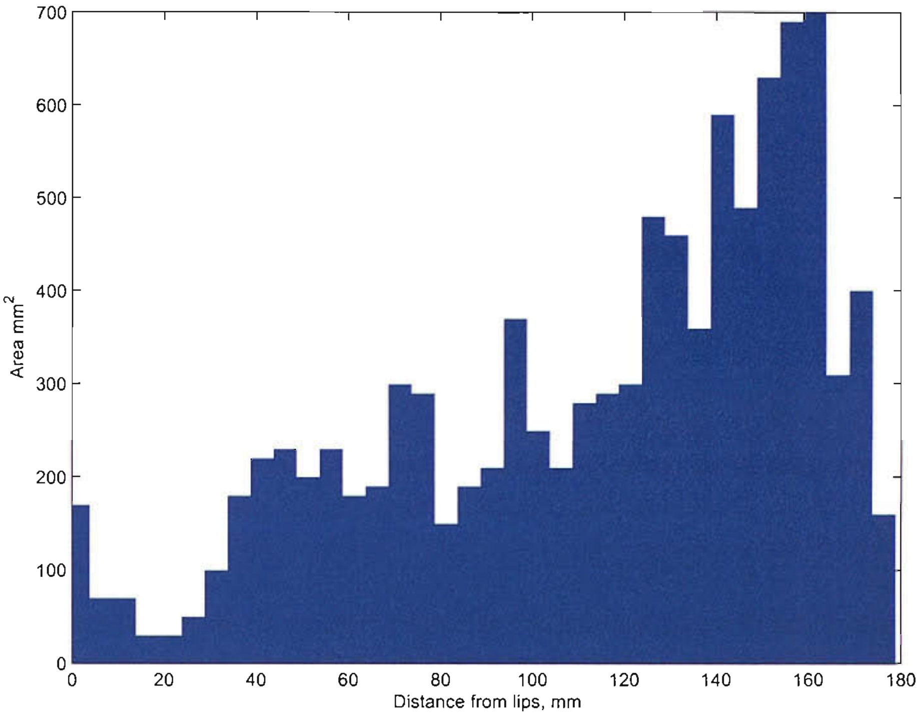

FIGURE A.5: Area function from Narayanan and Alwan (2000) of subject MI sustaining the fricative /s/. Data was copied from the published article.

the TFs of an area function derived from an actual vocal tract geometry.

A.1.3 Area Function of /s/

In this final test, ACTRA was used to calculated the transfer function for the area function of /s/ (copied from Narayanan and Alwan 2000). Their study focused on deriving an optimal noise source model for fricative modelling. Their area functions were derived from MRI data and the TFs were calculated using Maeda's model, a digital simulation software that is well established and frequently used in the field of speech synthesis (Maeda 1982).

The area function for the male subject, MI, sustaining /s/ is shown in Figure A.5. The constituent properties for the modelling were adopted from the article: reflecting walls with a resistance of 16000 kgs/m^2 and mass of 15 kg/m^2, speed of sound of 344.8 m/s and ratio of specific heats of 1.4. The far-field response of the TF was computed at a distance of 29 cm from the lips. Since the area functions from Narayanan and Alwan (2000) represent the whole vocal tract configuration, they are large in number and more complex in geometry. Therefore, D4ACTRA was used to assign the elements of ACTRA for each unit of the area function.

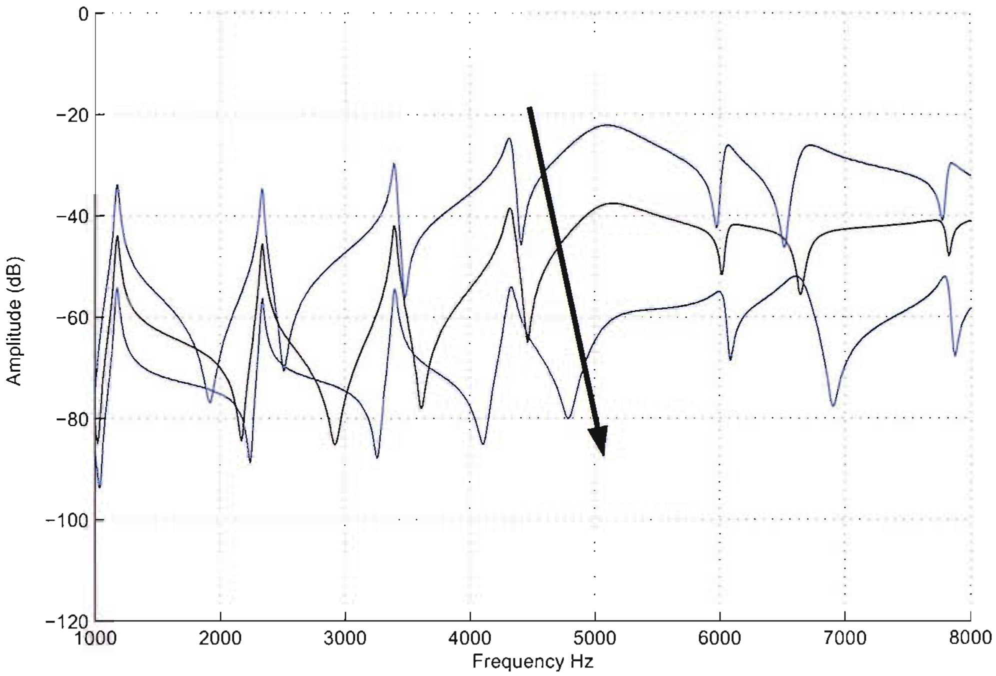

FIGURE A.6: Calculated TFs for area measurements of fricative /s/. The frequencies of the troughs are raised as the location of the pressure source is moved away from lips towards the constriction shown by the arrow. Refer to Figure 4(a) of Narayanan and Alwan (2000) for comparison.

The predicted spectra for this area function is shown in Figure A.6 for several pressure source locations. Even though the area function was copied from the article with much care, it must be emphasised that there is still the possibility of error through copying. The hydraulic radius parameters were also unavailable for the area function and as a result, the tract was assumed to have circular cross-sections throughout.

With these facts in mind, a comparison was made between the predicted spectra of ACTRA with the first five spectra of Figure 4 from Narayanan and Alwan (2000). Similarities between the spectra were observed in the following features: the broad peaks at approximately 5.5 kHz and the distinctive valleys above 2.0 kHz, which rises in frequency as the pressure source moves towards the constriction. The peaks in the vicinity of 3.0 and 4.5 kHz appear less defined in the Maeda spectra; nevertheless the frequencies seem to match well to the peaks in the spectra predicted by ACTRA. The spectra predicted by Maeda's model are obviously smoother—more like a spectral envelope, whereas the predicted spectra from ACTRA have very distinctive pole and zero frequencies. This may be caused by the distributed source that was employed in their modelling, whereas the ACTRA-predicted spectra assumed a single concentrated source. In general, it is possible to conclude that given the differences in the modelling, and the possibility of some error from the original area function while copying the data, the predicted spectra calculated by ACTRA show similarities to the distinct features of the spectra generated by Maeda's model.

A.1.4 Modelling Side Branches

The effects of side branches on the acoustic spectra of fricatives are relatively unknown, and it may be advantageous to model them realistically as separate ducts branching off the main airway. Frequently in the past, side branches of the vocal tract were simply modelled as an additional area to the main tract. This is acceptable provided that the new dimensions (i.e., after inclusion of the side cavity area) are still small compared to the wavelength of the sound (Ford 1970). However, experiments by Dang and Honda (1997) showed that for the vowels, pyriform sinuses should be realistically modelled as separate ducts, because the spectrum show incorrect behaviour in the frequency region (usually between 4.0–5.0 kHz) where the anti-resonance of the side branch is located. Therefore, modelling side cavities as an expansion of area may not be sufficient to model the system zeros correctly.

ACTRA is able to incorporate the areas of side cavities to some extent. The discrepancy lies in the fact that in an ideal vocal tract configuration, the area of the side cavity (specifically the pyriform sinuses) tapers off gradually like a cone, but in ACTRA, the user is able to specify only one cross-sectional area and length for the side branch. We have to assume that this model is a sufficient representation of the side branch. At least it provides a better approximation of the transfer function than an increased area, because it will introduce zeros (at the anti-resonances of the side branch) into the system.

The `outlet` element described earlier and shown in Figure 5.2 is used to model the side cavities of Figure A.7. The main airway of the tube was kept very short (~0.001 cm) to ensure that the resonances mainly reflected those of the side branch. The side duct has a length of 6.0 cm and a cross-sectional area of 4.0 cm^2. The main airway was also given a cross-sectional area of 4.0 cm^2. The walls were set to rigid and the properties of the medium were the same as the uniform tube model (see Section A.1.1).

A side branch that is a tube closed on one end will have a similar effect to a Helmholtz resonator; the acoustic impedance will be zero at all of the resonant frequencies of the tube (Ford 1970). The resonance of a Helmholtz resonator with no neck is given by:

$$F_H = \frac{c_0}{2\pi}\sqrt{\frac{2r}{V}}, \tag{A.4}$$

where r is the radius of branch opening (2.0 cm) and V is the volume of the cavity (Smith et al. 1996). Since the tube is a quarter-wave duct, the resonances will occur at F_H, $3F_H$, $5F_H$ and so on. The TF calculated by ACTRA for the model shown in Figure A.7 is shown in Figure A.8. The resonance frequencies are calculated using Equation A.4 and

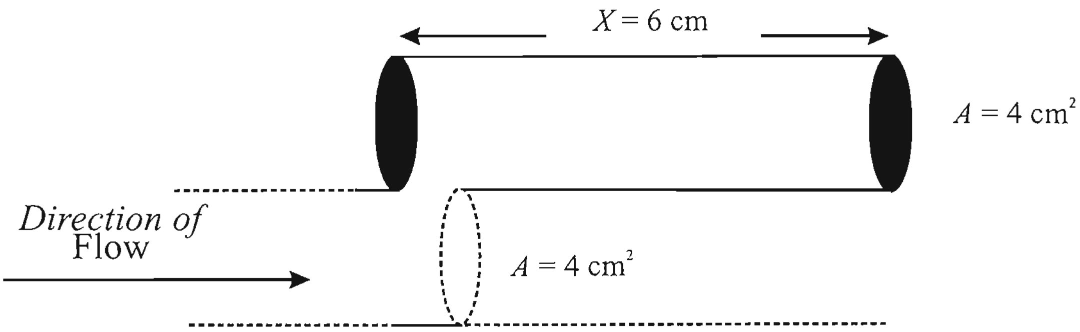

FIGURE A.7: Model of short tube with side branch. ACTRA models the side branch as an orifice.

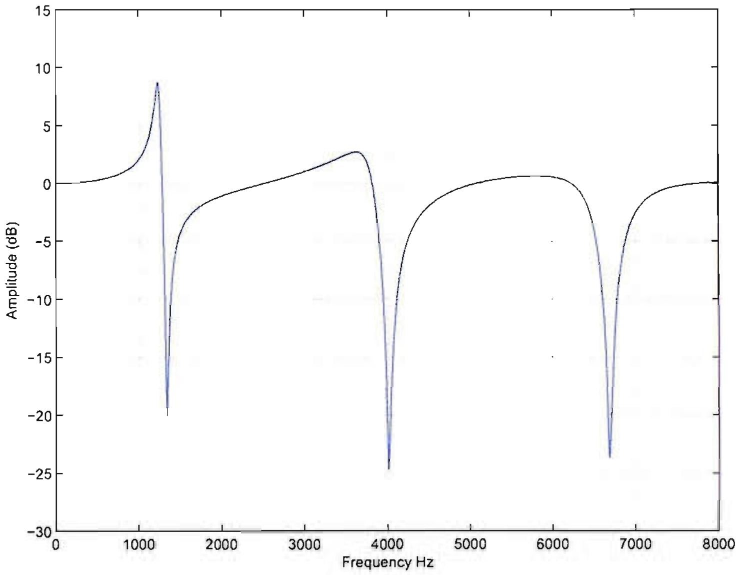

FIGURE A.8: Transfer function of simple tube with side branch.

System zeros	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
Calculated	1294	3882	6470
Predicted (ACTRA)	1340	4000	6690
% Error	3.5	3.0	3.4

TABLE A.2: Calculated and predicted zeros of model with side branch.

the zeros calculated by ACTRA are given in Table A.2. The calculated error between the zero frequencies was less than 3.5%.

A.2 Conclusions

The results of these tests allow us to confirm that the program is a reliable acoustic prediction software. The calculated error of approximately 4.0% overall indicates that it is capable of calculating TFs with an accuracy of up to 95%, as long as the assumption of 1-D wave propagation remains valid.

Appendix B

Details of Listening Experiment

This section includes the order of the presentation of sounds for each test in the listening experiment and the scores obtained/given by each listener.

Presentation	N_1	N_2	M	Presentation	N_1	N_2	M	Presentation	N_1	N_2	M	Presentation	N_1	N_2	M
1	/s/	/ʃ/	/s/	25	/θ/	/s/	/θ/	49	/ʃ/	/θ/	/ʃ/	73	/s/	/ʃ/	/ʃ/
2	/f/	/θ/	/f/	26	/ʃ/	/s/	/s/	50	/f/	/θ/	/f/	74	/ʃ/	/θ/	/θ/
3	/ʃ/	/f/	/ʃ/	27	/θ/	/f/	/f/	51	/ʃ/	/f/	/ʃ/	75	/θ/	/s/	/s/
4	/s/	/θ/	/θ/	28	/ʃ/	/s/	/ʃ/	52	/ʃ/	/s/	/s/	76	/f/	/θ/	/f/
5	/θ/	/ʃ/	/ʃ/	29	/f/	/s/	/s/	53	/f/	/θ/	/θ/	77	/s/	/ʃ/	/s/
6	/θ/	/f/	/θ/	30	/θ/	/ʃ/	/θ/	54	/ʃ/	/f/	/ʃ/	78	/s/	/f/	/f/
7	/s/	/f/	/f/	31	/ʃ/	/θ/	/ʃ/	55	/ʃ/	/s/	/ʃ/	79	/θ/	/f/	/θ/
8	/ʃ/	/θ/	/θ/	32	/θ/	/s/	/s/	56	/f/	/s/	/f/	80	/f/	/θ/	/θ/
9	/s/	/f/	/s/	33	/s/	/ʃ/	/s/	57	/s/	/θ/	/θ/	81	/ʃ/	/f/	/ʃ/
10	/f/	/ʃ/	/f/	34	/θ/	/f/	/θ/	58	/ʃ/	/f/	/f/	82	/θ/	/s/	/s/
11	/θ/	/ʃ/	/ʃ/	35	/f/	/s/	/s/	59	/s/	/ʃ/	/ʃ/	83	/θ/	/f/	/f/
12	/f/	/θ/	/θ/	36	/f/	/ʃ/	/f/	60	/ʃ/	/θ/	/θ/	84	/θ/	/ʃ/	/θ/
13	/ʃ/	/s/	/s/	37	/s/	/θ/	/s/	61	/f/	/ʃ/	/ʃ/	85	/ʃ/	/s/	/ʃ/
14	/s/	/f/	/f/	38	/f/	/ʃ/	/ʃ/	62	/ʃ/	/s/	/s/	86	/f/	/s/	/s/
15	/θ/	/ʃ/	/θ/	39	/θ/	/f/	/θ/	63	/f/	/s/	/f/	87	/θ/	/s/	/θ/
16	/ʃ/	/θ/	/ʃ/	40	/θ/	/f/	/f/	64	/s/	/f/	/s/	88	/ʃ/	/f/	/f/
17	/s/	/f/	/s/	41	/θ/	/ʃ/	/ʃ/	65	/ʃ/	/s/	/ʃ/	89	/s/	/ʃ/	/s/
18	/f/	/ʃ/	/ʃ/	42	/s/	/f/	/s/	66	/θ/	/ʃ/	/ʃ/	90	/s/	/θ/	/s/
19	/f/	/s/	/f/	43	/s/	/θ/	/θ/	67	/f/	/ʃ/	/f/	91	/s/	/θ/	/θ/
20	/s/	/ʃ/	/ʃ/	44	/θ/	/s/	/θ/	68	/s/	/θ/	/s/	92	/s/	/ʃ/	/ʃ/
21	/ʃ/	/f/	/f/	45	/ʃ/	/f/	/f/	69	/ʃ/	/θ/	/ʃ/	93	/f/	/s/	/f/
22	/f/	/θ/	/θ/	46	/s/	/f/	/f/	70	/f/	/s/	/s/	94	/f/	/ʃ/	/ʃ/
23	/f/	/θ/	/f/	47	/θ/	/s/	/s/	71	/θ/	/s/	/θ/	95	/ʃ/	/θ/	/θ/
24	/s/	/θ/	/s/	48	/θ/	/ʃ/	/θ/	72	/θ/	/f/	/f/	96	/f/	/ʃ/	/f/

TABLE B.1: Presentation of sounds for the MvN test. Each presentation consists of three sound segments. N_i and M denote the natural segments and the Mermelstein-derived segments respectively.

No	Subject	/f/*	/θ/	/s/	/ʃ/
1	ZH	8	3	6	7
2	AT	13	3	5	3
3	SC	3	6	8	7
4	BC	13	3	4	4
5	SB	10	5	2	7
6	DL	12	1	5	6
7	CB	6	5	6	7
8	JG	18	0	0	6
9	SP	8	4	4	8
10	BN	8	6	5	5
11	MN	8	6	4	6
12	BL	5	6	6	7
13	JD	11	4	4	4
14	SS	9	3	5	7
15	JP	11	4	5	4
16	AC	16	2	2	4
17	ES	12	4	5	3
18	AH	9	6	6	3

TABLE B.2: Results of 18 listeners for the Mermelstein-derived /f/ of the MvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/*	/s/	/ʃ/
1	ZH	4	19	1	0
2	AT	15	11	4	4
3	SC	2	10	5	7
4	BC	4	14	3	3
5	SB	3	10	5	6
6	DL	4	13	3	4
7	CB	2	12	7	3
8	JG	3	8	7	6
9	SP	6	7	5	6
10	BN	5	4	8	7
11	MN	4	12	3	5
12	BL	2	8	6	8
13	JD	4	5	8	7
14	SS	7	9	2	6
15	JP	4	10	4	6
16	AC	4	11	5	4
17	ES	7	10	5	2
18	AH	5	10	3	6

TABLE B.3: Results of 18 listeners for the Mermelstein-derived /θ/ of the MvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/	/s/*	/ʃ/
1	ZH	1	1	21	1
2	AT	2	1	18	3
3	SC	2	0	18	4
4	BC	0	0	24	0
5	SB	1	3	12	8
6	DL	4	3	15	2
7	CB	5	2	13	4
8	JG	4	2	10	8
9	SP	0	1	17	6
10	BN	2	3	19	0
11	MN	0	0	21	3
12	BL	3	0	14	7
13	JD	3	1	14	6
14	SS	2	1	16	5
15	JP	3	6	12	3
16	AC	3	2	15	4
17	ES	3	4	13	4
18	AH	0	0	21	3

TABLE B.4: Results of 18 listeners for the Mermelstein-derived /s/ of the MvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/	/s/	/ʃ/*
1	ZH	1	1	2	20
2	AT	2	0	5	17
3	SC	0	2	3	19
4	BC	1	2	2	19
5	SB	1	2	4	17
6	DL	3	2	2	17
7	CB	2	4	2	16
8	JG	0	0	0	24
9	SP	1	1	3	19
10	BN	0	1	0	23
11	MN	1	1	3	19
12	BL	1	1	0	22
13	JD	1	1	2	20
14	SS	1	0	2	21
15	JP	3	2	5	14
16	AC	4	3	4	13
17	ES	4	0	3	17
18	AH	1	0	3	20

TABLE B.5: Results of 18 listeners for the Mermelstein-derived /ʃ/ of the MvN test. The * denotes the number of times the listeners selected the correct category of sound.

Presentation	N_1	N_2	B	Presentation	N_1	N_2	B	Presentation	N_1	N_2	B	Presentation	N_1	N_2	B
1	/ʃ/	/s/	/ʃ/	25	/s/	/ʃ/	/s/	49	/s/	/f/	/s/	73	/θ/	/ʃ/	/ʃ/
2	/θ/	/ʃ/	/ʃ/	26	/s/	/θ/	/s/	50	/f/	/ʃ/	/ʃ/	74	/s/	/f/	/s/
3	/f/	/ʃ/	/f/	27	/s/	/θ/	/θ/	51	/f/	/s/	/f/	75	/s/	/θ/	/θ/
4	/s/	/θ/	/s/	28	/s/	/ʃ/	/ʃ/	52	/s/	/ʃ/	/ʃ/	76	/θ/	/s/	/θ/
5	/ʃ/	/θ/	/ʃ/	29	/f/	/s/	/f/	53	/ʃ/	/f/	/f/	77	/ʃ/	/f/	/f/
6	/f/	/s/	/s/	30	/f/	/ʃ/	/ʃ/	54	/f/	/θ/	/θ/	78	/s/	/f/	/f/
7	/θ/	/s/	/θ/	31	/ʃ/	/θ/	/θ/	55	/f/	/θ/	/f/	79	/θ/	/s/	/s/
8	/θ/	/f/	/f/	32	/f/	/ʃ/	/f/	56	/s/	/θ/	/s/	80	/θ/	/ʃ/	/θ/
9	/s/	/ʃ/	/ʃ/	33	/s/	/ʃ/	/s/	57	/θ/	/s/	/θ/	81	/ʃ/	/θ/	/ʃ/
10	/ʃ/	/θ/	/θ/	34	/f/	/θ/	/f/	58	/ʃ/	/s/	/s/	82	/f/	/θ/	/f/
11	/θ/	/s/	/s/	35	/ʃ/	/f/	/ʃ/	59	/θ/	/f/	/f/	83	/ʃ/	/f/	/ʃ/
12	/f/	/θ/	/f/	36	/s/	/θ/	/θ/	60	/ʃ/	/s/	/ʃ/	84	/ʃ/	/s/	/s/
13	/s/	/ʃ/	/s/	37	/θ/	/ʃ/	/ʃ/	61	/f/	/s/	/s/	85	/f/	/θ/	/θ/
14	/s/	/f/	/f/	38	/θ/	/f/	/θ/	62	/θ/	/ʃ/	/θ/	86	/ʃ/	/f/	/ʃ/
15	/θ/	/f/	/θ/	39	/s/	/f/	/f/	63	/ʃ/	/θ/	/ʃ/	87	/ʃ/	/s/	/ʃ/
16	/f/	/θ/	/θ/	40	/ʃ/	/θ/	/θ/	64	/θ/	/s/	/s/	88	/f/	/s/	/f/
17	/ʃ/	/f/	/ʃ/	41	/s/	/f/	/s/	65	/s/	/ʃ/	/s/	89	/s/	/θ/	/θ/
18	/θ/	/s/	/s/	42	/f/	/ʃ/	/f/	66	/θ/	/f/	/θ/	90	/ʃ/	/f/	/f/
19	/θ/	/f/	/f/	43	/θ/	/ʃ/	/ʃ/	67	/f/	/s/	/s/	91	/s/	/ʃ/	/ʃ/
20	/θ/	/ʃ/	/θ/	44	/f/	/θ/	/θ/	68	/f/	/ʃ/	/f/	92	/ʃ/	/θ/	/θ/
21	/ʃ/	/s/	/ʃ/	45	/ʃ/	/s/	/s/	69	/s/	/θ/	/s/	93	/f/	/ʃ/	/ʃ/
22	/f/	/s/	/s/	46	/s/	/f/	/f/	70	/f/	/ʃ/	/ʃ/	94	/ʃ/	/s/	/s/
23	/θ/	/s/	/θ/	47	/θ/	/ʃ/	/θ/	71	/θ/	/f/	/θ/	95	/f/	/s/	/f/
24	/ʃ/	/f/	/f/	48	/ʃ/	/θ/	/ʃ/	72	/θ/	/f/	/f/	96	/s/	/f/	/s/

TABLE B.6: Presentation of sounds for the BvN test. Each presentation consists of three sound segments. N_i and B denote the natural segments and the Blum-derived segments respectively.

No	Subject	/f/*	/θ/	/s/	/ʃ/
1	ZH	13	5	3	3
2	AT	11	5	4	4
3	SC	8	5	5	6
4	BC	16	3	1	4
5	SB	16	3	3	2
6	DL	15	3	4	2
7	CB	12	4	4	4
8	JG	8	5	5	6
9	SP	12	3	5	4
10	BN	16	2	4	2
11	MN	11	6	4	3
12	BL	12	4	5	3
13	JD	8	4	6	6
14	SS	15	4	4	1
15	JP	12	4	4	4
16	AC	14	5	1	4
17	ES	13	5	3	3
18	AH	15	2	4	3

TABLE B.7: Results of 18 listeners for the Blum-derived /f/ of the BvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/*	/s/	/ʃ/
1	ZH	4	16	2	2
2	AT	3	11	4	6
3	SC	2	10	6	6
4	BC	4	10	4	6
5	SB	4	10	5	5
6	DL	4	11	4	5
7	CB	2	16	3	3
8	JG	4	5	8	7
9	SP	4	13	3	4
10	BN	3	9	6	6
11	MN	5	11	4	4
12	BL	3	11	5	5
13	JD	5	5	8	6
14	SS	4	10	3	7
15	JP	4	12	4	4
16	AC	4	13	4	3
17	ES	2	14	3	5
18	AH	15	11	2	6

TABLE B.8: Results of 18 listeners for the Blum-derived /θ/ of the BvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/	/s/*	/ʃ/
1	ZH	0	2	19	3
2	AT	1	5	11	7
3	SC	2	2	12	8
4	BC	1	0	19	4
5	SB	1	6	13	4
6	DL	5	6	10	3
7	CB	2	2	18	2
8	JG	1	1	14	8
9	SP	5	4	9	6
10	BN	2	3	12	7
11	MN	5	2	14	3
12	BL	1	3	15	5
13	JD	2	3	16	3
14	SS	3	3	15	3
15	JP	3	2	15	4
16	AC	3	5	13	3
17	ES	7	5	7	5
18	AH	1	2	17	4

TABLE B.9: Results of 18 listeners for the Blum-derived /s/ of the BvN test. The * denotes the number of times the listeners selected the correct category of sound.

No	Subject	/f/	/θ/	/s/	/ʃ/*
1	ZH	2	1	1	20
2	AT	1	2	1	20
3	SC	1	0	2	21
4	BC	2	3	1	18
5	SB	2	2	2	18
6	DL	2	5	4	13
7	CB	3	2	2	17
8	JG	1	2	1	20
9	SP	6	2	3	13
10	BN	1	1	1	21
11	MN	3	4	2	15
12	BL	1	0	1	22
13	JD	2	2	1	19
14	SS	1	2	4	17
15	JP	2	1	4	17
16	AC	4	3	6	11
17	ES	7	5	4	8
18	AH	2	2	3	17

TABLE B.10: Results of 18 listeners for the Blum-derived /ʃ/ of the BvN test. The * denotes the number of times the listeners selected the correct category of sound.

Presentation	S_1	S_2	N	Presentation	S_1	S_2	N
1	B	M	/f/	21	M	B	/ʃ/
2	B	M	/ʃ/	22	M	B	/s/
3	M	B	/ð/	23	M	B	/f/
4	M	B	/s/	24	B	M	/ð/
5	B	M	/ð/	25	B	M	/ʃ/
6	B	M	/s/	26	B	M	/f/
7	M	B	/f/	27	B	M	/ʃ/
8	M	B	/s/	28	M	B	/s/
9	B	M	/ð/	29	B	M	/f/
10	B	M	/s/	30	B	M	/ð/
11	M	B	/f/	31	M	B	/ʃ/
12	B	M	/s/	32	M	B	/ð/
13	B	M	/ʃ/	33	B	M	/s/
14	B	M	/f/	34	B	M	/f/
15	M	B	/ð/	35	B	M	/ʃ/
16	B	M	/s/	36	B	M	/ð/
17	M	B	/ʃ/	37	M	B	/ʃ/
18	M	B	/ð/	38	M	B	/f/
19	M	B	/ʃ/	39	M	B	/ð/
20	M	B	/f/	40	M	B	/s/

TABLE B.11: Presentation of sounds for the MvB test. Each presentation consists of three sound segments. S_i, M, B and N denote the synthetic segments, the Mermelstein-derived segments, the Blum-derived segments and the natural segments respectively.

		/f/		/θ/		/s/		/ʃ/	
No	Subject	M	B	M	B	M	B	M	B
1	ZH	3	7	7	3	8	2	5	5
2	AT	5	5	8	2	4	6	10	0
3	SC	1	9	2	8	5	5	8	2
4	BC	5	5	4	6	9	1	8	2
5	SB	6	4	5	5	5	5	7	3
6	DL	5	5	5	5	5	5	2	8
7	CB	2	8	1	9	2	8	5	5
8	JG	4	6	8	2	8	2	8	2
9	SP	3	7	4	6	5	5	6	4
10	BN	4	6	7	3	7	3	10	0
11	MN	8	2	7	3	7	3	9	1
12	BL	3	7	4	6	3	7	5	5
13	JD	4	6	7	3	3	7	7	3
14	SS	3	7	3	7	7	3	4	6
15	JP	4	6	5	5	8	2	5	5
16	AC	5	5	5	5	5	5	5	5
17	ES	7	3	7	3	6	4	3	7
18	AH	2	8	7	3	6	4	4	6

TABLE B.12: Results of 18 listeners for the MvB test listed according to fricative. M and B denotes the Mermelstein- and Blum-derived versions respectively.

Presentation	S$_1$	S$_2$	N	Presentation	S$_1$	S$_2$	N
1	B	B$^+$	/s/	29	B*	B	/(i)f/
2	B*	B	/(a)f/	30	B**	B	/ʃ/
3	B*	B	/ʃ/	31	B**	B	/(a)f/
4	B*	B	/(i)f/	32	B	B*	/(i)f/
5	B	B**	/ʃ/	33	B	B$^+$	/s/
6	B	B*	/(a)f/	34	B	B**	/(a)f/
7	B	B**	/(i)f/	35	B**	B	/(i)f/
8	B	B**	/(a)f/	36	B	B**	/ʃ/
9	B**	B	/(i)f/	37	B**	B	/(a)f/
10	B	B*	/ʃ/	38	B	B*	/(i)f/
11	B	B**	/(a)f/	39	B	B**	/(a)f/
12	B	B*	/(i)f/	40	B**	B	/ʃ/
13	B$^+$	B	/s/	41	B	B$^+$	/s/
14	B*	B	/ʃ/	42	B	B**	/(i)f/
15	B*	B	/(a)f/	43	B	B*	/(a)f/
16	B	B$^+$	/s/	44	B	B*	/ʃ/
17	B*	B	/(i)f/	45	B*	B	/(i)f/
18	B**	B	/ʃ/	46	B*	B	/(a)f/
19	B$^+$	B	/s/	47	B**	B	/ʃ/
20	B	B*	/ʃ/	48	B$^+$	B	/s/
21	B	B*	/(a)f/	49	B	B*	/(i)f/
22	B*	B	/ʃ/	50	B	B**	/(i)f/
23	B	B**	/(i)f/	51	B**	B	/(a)f/
24	B	B*	/ʃ/	52	B*	B	/ʃ/
25	B**	B	/(a)f/	53	B**	B	/(i)f/
26	B**	B	/(i)f/	54	B	B**	/ʃ/
27	B$^+$	B	/s/	55	B*	B	/(a)f/
28	B	B*	/(a)f/	56	B	B**	/ʃ/

TABLE B.13: Presentation of sounds for the BvNB test. Each presentation consists of three sound segments. S$_i$ and N denote the synthetic Blum-derived segments and the natural respectively. B, B$^+$, B* and B** denote Blum-derived segments with no branches, with inclusion of the sublingual cavity, inclusion of the pyriform sinuses and inclusion of both respectively.

No	Subjects	/s/$^+$	/ʃ/*	/ʃ/**	/(a)f/*	/(a)f/**	/(i)f/*	/(i)f/**
1	ZH	6	3	3	4	4	3	4
2	AT	6	1	0	4	4	6	6
3	SC	5	2	3	4	5	6	5
4	BC	4	1	0	3	2	7	6
5	SB	5	0	0	4	1	3	5
6	DL	3	2	3	3	6	5	7
7	CB	2	0	0	6	4	5	8
8	JG	2	0	0	3	2	4	2
9	SP	5	3	3	6	5	6	3
10	BN	1	2	4	4	4	0	1
11	MN	6	0	0	4	6	3	2
12	BL	3	1	2	4	3	6	7
13	JD	5	1	0	7	7	8	8
14	SS	2	2	3	5	5	4	5
15	JP	3	3	5	5	2	4	5
16	AC	3	3	4	2	4	4	5
17	ES	5	4	3	4	4	4	4
18	AH	4	0	1	3	5	5	6

TABLE B.14: Results of 17 listeners for the models with side branches in the BvNB test. The values represent the number of times the listeners chose a model with a side branch over a model without side branches out of a total of eight times. The $^+$, *, and ** denote inclusion of the pyriform sinuses, the sublingual cavity and both respectively.

Presentation	Segment	Presentation	Segment	Presentation	Segment	Presentation	Segment	Presentation	Segment
1	B^{+} /s/	26	N /s/	51	B** /(a)f/	76	B** /(i)f/	101	B /θ/
2	M /(i)f/	27	B /(a)f/	52	B* /ʃ/	77	B /θ/	102	B* /ʃ/
3	B** /(a)f/	28	B^{+} /s/	53	B /θ/	78	N /(u)f/	103	N /(u)f/
4	B /(a)f/	29	B* /ʃ/	54	N /s/	79	M /s/	104	B /(u)f/
5	B* /(a)f/	30	B** /ʃ/	55	B** /ʃ/	80	M /θ/	105	B* /(a)f/
6	B /(i)f/	31	M /(u)f/	56	N /(u)f/	81	B /(a)f/	106	M /(u)f/
7	N /s/	32	N /(u)f/	57	M /s/	82	B /(u)f/	107	M /(i)f/
8	B /s/	33	B* /(i)f/	58	B /(a)f/	83	B /s/	108	B* /(i)f/
9	B** /ʃ/	34	B** /(a)f/	59	M /θ/	84	B* /ʃ/	109	N /(i)f/
10	M /(u)f/	35	N /(i)f/	60	B** /(i)f/	85	N /ʃ/	110	B /ʃ/
11	B* /(i)f/	36	B /s/	61	M /ʃ/	86	M /(i)f/	111	M /ʃ/
12	maf	37	B /(i)f/	62	B /(u)f/	87	M /ʃ/	112	N /ʃ/
13	N /ʃ/	38	M /(i)f/	63	M /(i)f/	88	B* /(i)f/	113	B** /(i)f/
14	N /(u)f/	39	M /ʃ/	64	B* /(i)f/	89	B* /(a)f/	114	B^{+} /s/
15	B /ʃ/	40	B /(u)f/	65	N /ʃ/	90	B^{+} /s/	115	N /θ/
16	B* /ʃ/	41	N /ʃ/	66	M /(u)f/	91	M /(u)f/	116	B /s/
17	B /(u)f/	42	N /θ/	67	N /(i)f/	92	B /ʃ/	117	B /(i)f/
18	M /ʃ/	43	B /θ/	68	B* /(a)f/	93	N /θ/	118	B** /(a)f/
19	N /θ/	44	B* /(a)f/	69	N /θ/	94	N /s/	119	B /(a)f/
20	B /θ/	45	B /(a)f/	70	B /ʃ/	95	B** /ʃ/	120	M /s/
21	N /(i)f/	46	B** /(i)f/	71	B /s/	96	N /(i)f/	121	M /θ/
22	B /(a)f/	47	maf	72	B /(i)f/	97	B** /(a)f/	122	maf
23	M /θ/	48	B /ʃ/	73	maf	98	B /(i)f/	123	B** /ʃ/
24	B** /(i)f/	49	M /s/	74	B /(a)f/	99	maf	124	B /(a)f/
25	M /s/	50	M /θ/	75	B^{+} /s/	100	B /(a)f/	125	N /s/

TABLE B.15: Presentation of sounds for the RtS test. Each presentation consists of one sound segment and and information on their identity (i.e., /f/ as in *f*un, /s/ as in *s*un, /th/ as in *th*orn, /sh/ as in *sh*ine. N, M, and B denote the natural segments, the Mermelstein-derived segments and the Blum-derived segments (without side branches) respectively. B^{+}, B* and B** denotes the Blum-derived segments with inclusion of the pyriform sinuses, sublingual cavity and both respectively.

		/(a)f/			/(i)f/			/(u)f/		
No	Subject	N	M	B	N	M	B	N	M	B
1	ZH	1.6	1.8	1.8	1.6	2.0	1.8	2.0	1.6	2.4
2	AT	1.6	1.4	1.6	2.0	2.2	2.8	1.6	3.2	2.8
3	SC	2.6	2.8	2.8	2.4	1.0	1.2	3.0	2.6	1.8
4	BC	3.2	1.4	1.8	3.0	1.8	2.0	3.2	2.2	2.0
5	SB	1.6	3.0	2.2	2.4	2.4	2.8	3.0	3.0	2.4
6	DL	2.6	1.2	1.4	2.0	1.8	2.0	2.4	2.0	1.0
7	CB	3.0	1.6	2.0	3.2	1.2	1.2	3.4	1.8	1.0
8	JG	1.0	1.2	1.0	1.2	1.4	1.6	1.2	1.0	1.4
9	SP	2.6	3.0	2.8	4.0	1.6	1.0	4.2	2.4	1.2
10	BN	1.8	3.2	2.6	2.2	1.2	2.2	2.4	2.8	2.0
11	MN	1.2	1.4	1.0	1.0	1.2	1.2	1.0	1.2	1.0
12	BL	1.2	2.8	3.2	2.2	2.4	2.2	1.6	3.2	1.8
13	JD	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.2
14	SS	2.2	2.4	1.4	1.6	2.8	2.2	2.2	2.8	2.4
15	JP	1.4	1.4	1.4	1.2	1.0	1.4	1.6	1.6	1.4
16	AC	2.0	2.4	2.2	2.2	1.8	1.6	2.6	2.4	1.4
17	ES	2.4	2.0	2.6	2.4	3.0	2.2	2.6	2.6	2.6
18	AH	3.0	2.0	3.2	2.6	3.0	3.2	2.4	3.0	3.8

TABLE B.16: The ratings of the RtS test given by 18 listeners for /f/ in /a, i, u/ context. N, M and B denote the natural segments, the Mermelstein-derived segments and the Blum-derived segments respectively.

		/θ/			/s/			/ʃ/		
No	Subject	N	M	B	N	M	B	N	M	B
1	ZH	2.6	1.6	1.8	3.8	4.4	4.2	4.2	4.2	4.8
2	AT	2.6	3.8	3.4	1.8	3.2	4.0	3.2	3.6	3.8
3	SC	1.6	1.2	1.0	2.2	2.2	2.4	3.4	4.0	3.8
4	BC	2.2	2.2	2.4	2.6	3.8	4.0	3.6	4.2	4.2
5	SB	2.0	1.6	1.4	2.2	2.6	2.4	3.2	4.4	4.2
6	DL	3.2	2.8	1.4	3.2	2.8	4.0	3.8	4.0	4.2
7	CB	1.4	1.6	1.4	2.2	2.0	2.8	4.4	4.4	4.2
8	JG	1.0	2.6	2.4	2.0	2.8	3.0	2.4	3.0	2.8
9	SP	1.8	1.2	1.4	3.4	2.4	2.8	3.8	3.4	3.6
10	BN	1.2	2.0	1.0	2.0	2.2	2.0	3.0	3.6	4.0
11	MN	1.4	1.4	1.2	2.2	2.4	2.6	2.2	2.0	1.8
12	BL	2.4	2.8	2.2	2.0	2.8	2.8	3.2	4.6	4.2
13	JD	1.0	1.2	1.6	1.4	2.0	2.0	2.0	2.0	2.4
14	SS	2.2	2.6	1.8	3.4	3.6	3.4	4.4	4.0	4.6
15	JP	1.2	1.4	1.2	3.4	3.8	3.4	4.8	4.8	4.6
16	AC	2.4	2.8	1.6	2.8	3.0	2.8	4.0	4.6	5.0
17	ES	2.2	2.4	2.2	2.6	2.8	2.8	3.8	3.8	4.0
18	AH	2.4	3.6	3.2	2.6	3.6	2.8	3.2	3.2	3.2

TABLE B.17: The ratings of the RtS test given by 18 listeners for /θ, s,ʃ/. N, M and B denote the natural segments, the Mermelstein-derived segments and the Blum-derived segments respectively.

		/s/	/ʃ/		/(a)f/		/(i)f/	
No	Subject	B^+	B*	B**	B*	B**	B*	B**
1	ZH	4.0	4.2	4.6	2.0	2.2	2.0	2.0
2	AT	3.6	3.6	3.8	2.0	2.2	2.6	2.8
3	SC	2.4	3.8	3.6	2.6	2.4	1.4	1.2
4	BC	3.0	4.4	4.4	2.0	1.4	1.8	1.8
5	SB	2.0	3.8	4.2	2.0	2.4	2.2	2.2
6	DL	3.6	3.6	4.6	2.2	1.6	1.8	2.4
7	CB	3.0	4.2	4.0	2.6	1.8	1.0	1.0
8	JG	2.8	3.6	3.2	1.2	1.0	1.2	1.6
9	SP	2.6	3.0	3.2	2.2	2.8	1.2	2.0
10	BN	2.4	3.2	3.4	2.2	2.4	1.8	1.6
11	MN	2.4	2.0	2.4	1.2	1.2	1.2	1.0
12	BL	3.0	3.6	3.8	3.0	2.6	1.8	1.8
13	JD	2.4	1.8	1.8	1.0	1.0	1.0	1.0
14	SS	3.6	4.6	4.4	1.8	2.2	2.8	2.8
15	JP	3.8	4.2	4.6	1.6	1.4	1.2	1.2
16	AC	2.8	3.8	4.2	2.2	2.4	1.6	1.6
17	ES	2.2	3.4	4.2	2.2	2.6	3.0	2.2
18	AH	2.8	3.0	4.0	3.4	3.2	3.4	2.8

TABLE B.18: The ratings of the RtS test given by 18 listeners for /s, ʃ, (a)f, (i)f/ derived through the Blum technique with inclusion of the side branches. B^+, B* and B** denotes inclusion of the pyriform sinuses, sublingual cavity and both respectively.

Bibliography

Badin, P. (1989). Acoustic of voiceless fricatives: Production theory and data. Technical report, Department of Speech, Music and Hearing, `http://www.speech.kth.se` and `http://www.speech.kth.se/qpsr`. Speech Transmission Laboratory – Quarterly Progress and Status Report.

Badin, P., G. Bailly, M. Raybaudi, and C. Segebarth (1998). A three-dimensional linear articulatory model based on MRI data. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 2, Sydney, Australia, pp. 417–420.

Badin, P., I. S. Makarov, and V. N. Sorokin (2005). Algorithm for calculating the cross-section areas of the vocal tract. *Acoustical Physics 51*(1), 52–58.

Baer, T., J. C. Gore, L. C. Graco, 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.

Beautemps, D., P. Badin, and R. Laboissière (1995). Deriving vocal tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. *Speech Communication 16*(1), 27–47.

Bendat, J. and A. Piersol (1993). *Engineering Applications of Correlation and Spectral Analysis.* John Wiley, New York, NY.

Beranek, L. L. (1949). *Acoustic Measurements.* John Wiley, New York, NY.

Best, C. T., B. Morrongiello, and R. Robson (1981). Perceptual equivalence of acoustic cues in speech and non-speech perception. *Perception and Psychophysics 29*, 191.

Blum, H. (1973). Biological shape and visual science: Part 1. *Journal of Theoretical Biology 38*, 205–287.

Catford, J. C. (1977). *Fundamental Problems in Phonetics.* Indiana University Press, Bloomington, IN.

Chomsky, N. and M. Halle (1968). *The Sound Patterns of English.* MIT Press, Cambridge, MA.

Coker, C. and O. Fujimura (1966). Model for specification of the vocal tract area function. *Journal of the Acoustical Society of America 40*, 1271. (abs).

Coker, C. H. (1968). Speech synthesis with a parametric articulatory model. In *Proceedings of the Speech Symposium*, Kyoto, Japan. Paper A4.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE 64*(4), 452–460.

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., K. Honda, and H. Suzuki (1993). MRI measurements and acoustic investigation of the nasal and para-nasal cavities. *Journal of the Acoustical Society of America 94*, 1765. (abs).

Davies, P. O. A. L. (1988). Practical flow duct acoustics. *Journal of Sound and Vibration 124*(1), 91–115.

Davies, P. O. A. L., R. S. McGowan, and C. H. Shadle (1993). Practical flow duct acoustics applied to the vocal tract. In R. Titze (Ed.), *Vocal Fold Physiology: Frontiers in Basic Science*, pp. 93–142. Singular Publishers, San Diego, CA.

Demolin, D., T. Metens, and A. Soquet (1996). Three-dimensional measurement of the vocal tract by MRI. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 1, PA, pp. 272–275.

Engwall, O. (1992). A revisit to the application of MRI to the analysis of speech production – testing our assumptions. In *Proceedings of the 6th International Seminar on Speech Production*, Volume 35, Sydney, Australia, pp. 53–67.

Engwall, O. and P. Badin (1999). Collecting and analysing two- and three-dimensional MRI data for Swedish. In *Speech, Music and Hearing (TMH) – Quarterly Progress and Status Report*, Volume 4, pp. 1–28.

Engwall, O. and P. Badin (2000). An MRI study of Swedish fricatives: Coarticulatory effects. In *Proceedings of the 5th International Seminar on Speech Production*, Kloster Seeon, Germany, pp. 297–300.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton De Gruyter, the Netherlands.

Fant, G. and M. Båveguård (1995). Parametric model of VT area functions: Vowels and consonants. *EU Report, Speechmaps*. Delivery 28, WP2.2, 1–30.

Flanagan, J. L. (1972). *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, Germany.

Foldvik, A. K., O. Husby, J. Kværness, I. C. Nordli, and P. A. Rinck (1990). MRI film of articulatory movements. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 1, Kobe, Japan, pp. 421–422.

Ford, R. D. (1970). *Introduction to Acoustics*. Elsevier, Great Yarmouth, UK.

Gibbs, J. and E. Wilson (1960). *Vector Analysis: A Text-Book for the use of Students of Mathematics and Physics, Founded Upon the Lectures of J. Willard Gibbs.* Dover, New York, NY.

Goldstein, U. G. (1980). *An Articulatory Model for the Vocal Tracts of Growing Children.* Ph. D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Greenwood, A. R., C. C. Goodyear, and P. A. Martin (1992). Measurements of vocal tract shapes using magnetic resonance imaging. *IEEE Proceedings -I 139*(6), 553–560.

Harris, K. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech 1*, 1–7.

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 (1964). On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *Journal of the Acoustical Society of America 36*, 1037.

Heinz, J. M. and K. N. Stevens (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proceedings of the 5th International Congress on Acoustics*, Liège, Belgium, pp. 1–4. Paper A44.

Henke, W. L. (1966). *Dynamic Articulatory Model of Speech Production using Computer Simulation.* Ph. D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Holtrup, G. (1998). From Magnetic Resonance Imaging (MRI) Data to Simulation of Speech Sounds. Technical report, School of Electronics and Computer Science, University of Southampton, Southampton, UK. Tripartite Fifth Year Project Report.

Ishizaka, K., J. C. French, and J. L. Flanagan (1975). Direct determination of vocal tract wall impedance. *IEEE Transactions on Acoustics, Speech and Signal Processing 23*, 370–373.

Jackson, P. J. B. (2000). *Characterisation of Plosive, Fricative and Aspiration Components in Speech Production.* Ph. D. thesis, School of Electronics and Computers Science, University of Southampton, Southampton, UK.

Johnson, K. and J. Ralston (1994). Automaticity in speech perception: some speech/nonspeech comparisons. *Phonetica 51*(4), 195–209.

Johnson, R. R. (1988). *Elementary Statistics* (5th ed.). PWS-KENT, Boston, MA.

Kinsler, L. E., A. R. Frey, A. B. Coppens, and J. V. Sanders (1982). *Fundamentals of Acoustics* (3rd ed.). John Wiley, New York, NY.

Ladefoged, P., J. Anthony, and D. Riley (1971). Direct measurements of the vocal tract. *Journal of the Acoustical Society of America 49*, 104.

Lighthill, M. J. (1954). On sound generated aerodynamically: II. Turbulence as a source of sound. *Proceedings of the Royal Society 1*, A222.

Lin, Q. (1990). *Speech Production Theory and Articulatory Speech Synthesis.* Ph. D. thesis, Royal Institute of Technology (KTH), Stockholm, Sweden.

Maeda, S. (1972). On the conversion of vocal tract X-ray data into formant frequencies. Technical report, Bell Laboratories, Murray Hill, NJ.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication 1*, 199–229.

Mermelstein, P. (1967). On the piriform recesses and their acoustic effects. *Folia Phoniatr 19*(5), 388–389.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America 53*(4), 1070–1082.

Mermelstein, P., S. Maeda, and O. Fujimura (1971). Description of the tongue lip movement in a jaw-based coordinate system. *Journal of the Acoustical Society of America 49*, 104. (abs).

Miller, J. E. and O. Fujimura (1975). From tongue model data to sound. *Journal of the Acoustical Society of America 57*, S3. (abs).

Moore, C. A. (1992). The correspondence of vocal tract images with volumes obtained from magnetic resonance images. *Journal of Speech and Hearing Research 35*, 1009–1023.

Munhall, K., E. Vatikiotis-Bateson, and Y. Tohkura (1994). *X-ray Database for Speech Research.* `http://pavlov.psyc.queensu.ca/faculty/munhall/x-ray/`, Queen's University, Kingston, Canada.

Narayanan, S. S. and A. A. Alwan (2000). Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing 8*, 328–344.

Narayanan, S. S., A. A. Alwan, and K. Haker (1995). Articulatory study of fricative consonants using Magnetic Resonance Imaging. *Journal of the Acoustical Society of America 3*, 1325–1347.

Quistgaard, J. U. (1997, Jan). Signal acquisition and processing in medical diagnostic ultrasound. *IEEE Signal Processing Magazine 14*(1), 67–74.

Rabiner, L. R. and R. W. Schafer (1978). *Digital Processing of Speech Signals.* Signal Processing Series. Prentice Hall, Englewood Cliffs, New Jersey.

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., E. Grabe-Georges, and E. Castelli (1992). Articulatory paths for some fricatives in connected speech. *Speech Communication 11*, 411–416.

Shadle, C. H. (1985). The acoustics of fricative consonants. Technical report, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Shadle, C. H. (1990). Articulatory-acoustic relationships in fricative consonants. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 187–209. Kluwer, the Netherlands.

Shadle, C. H. (1991). The effect of geometry on source mechanisms. *Journal of Phonetics 19*, 409–424.

Shadle, C. H., P. Badin, and A. Moulinier (1991). Towards the spectral characteristics of fricative consonants. In *Proceedings of the International Congress of Phonetic Science*, Volume 3, Aix-en-Provence, France, pp. 42–45.

Shadle, C. H., J. N. Carter, and S. J. Mair (1996). Acoustic characteristics of the front fricatives [f, v, θ, ð]. In *Proceedings of the 1st ETRW and the 4th International Seminar on Speech Production*, Autrans, France, pp. 193–196.

Shadle, C. H., S. J. Mair, and J. N. Carter (1995). The effect of vowel context on acoustic characteristics of [ç,x]. In *Proceedings of the International Congress of Phonetic Science*, Volume 1, Stockholm, Sweden, pp. 66–69.

Shadle, C. H., A. Moulinier, C. Dobelke, and C. Scully (1992). Ensemble averaging applied to the analysis of fricative consonants. In *Proceedings of the International Conference on Spoken Language Processing*, Volume 1, Banff, Canada, pp. 53–56.

Shadle, C. H. and C. Scully (1995). An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *Journal of Phonetics 23*, 53–66.

Shadle, C. H., M. Tiede, S. Masaki, Y. Shimada, and I. Fujimoto (1996). An MRI study of the effects of vowel context on fricatives. *Proceedings of the Institute of Acoustics 18*(9), 187–193.

Smith, B. J., R. J. Peters, and S. Owen (1996). *Acoustics and Noise Control* (2nd ed.). Longman Group, UK.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America 50*(4), 1180–1192. pt. 2.

Stevens, K. N. (1985). Evidence for the role of acoustic boundaries in the perception of speech sounds. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, Chapter 16, pp. 243–255. Academic Press, New York, NY.

Stevens, K. N. (1998). *Acoustic Phonetics*. Massachusetts Institute of Technology Press, Cambridge, MA.

Stone, M. and A. Lundberg (1996). Three dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America 99*(6), 3728–3737.

Stone, M., D. Ong, and A. Lundberg (1996). An MRI and EPG examination of /r/ and /l/. *Journal of the Acoustical Society of America 100*(4), 2660. (abs).

Story, B., I. Titze, and E. Hoffman (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America 100*(1), 537–554.

Subari, K. S., C. H. Shadle, A. Barney, and R. I. Damper (2004). Comparison of fricative vocal tract transfer functions derived using two different segmentation techniques. In *Proceedings of the International Conference of Signal Processing*, Istanbul, Turkey, pp. 164–168.

Sulter, A. M., D. G. Miller, R. F. Wolf, H. K. Schutte, H. P. Wit, and E. L. Mooyaart (1992). On the relation between the dimensions and resonance characteristics of the vocal tract: A study with MRI. *Magnetic Resonance Imaging 10*, 365–373.

Sundberg, J. (1974). Articulatory interpretation of the singing formants. *Journal of the Acoustical Society of America 55*, 838–844.

Tiede, M. K., S. Masaki, and E. Vatikiotis-Bateson (2000). Contrasts in speech articulation observed in sitting and supine conditions. In *Proceedings of the 5th International Seminar on Speech Production*, Kloster Seeon, Germany, pp. 25–28.

Tiede, M. K. and H. Yehia (1997). A parametric three-dimensional model of the vocal tract based on MRI data. In *International Conference on Acoustics, Speech, and Signal Processing*, Volume 3, Munich, Germany.

Whalen, D. H. (1991). Perception of the English [s]-[ʃ] distinction relies on fricative noises and transitions, not on brief spectral slices. *Journal of the Acoustical Society of America 90*(4), 1776–1785.

Wright, G. A. (1997, Jan). Magnetic resonance imaging. *IEEE Signal Processing Magazine 14*, 56–66.

Zhang, Z., S. Boyce, C. Espy-Wilson, and M. Tiede (2003). Acoustic strategies for production of American English “retroflex” /r/. In *Proceedings of the International Congress of Phonetic Science*, Volume 1, Barcelona, Span, pp. 1125–1128.