UNIVERSITY OF SOUTHAMPTON

# Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods

by

Oliver S. Blacklock

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

November 2004

UNIVERSITY OF SOUTHAMPTON

## ABSTRACT

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

Doctor of Philosophy

by Oliver S. Blacklock

An investigation into the properties of fricative production in normal and disordered speech is described. Methods for analysing and characterising fricative productions from acoustical signals can help provide better knowledge of the fricative production mechanisms. Being able to measure changes in the acoustical signal that relate to changes in production is extremely useful for the analysis of speech production in general, and for assessing long-term effects on speech of aids such as cochlear implants.

Parametric analysis based on mathematical models of the noise source and filter function of the tract (e.g. spectral tilt, and pole and zero frequencies) has been able to explain the behaviour of some elements of fricative production. However, not all aspects of fricative production can be accounted for by such models. Distinguishing characteristics and ranges of variation of all the fricatives have not been satisfactorily captured. The turbulent noise sources that are generated near constrictions within the tract behave in complex ways that cannot be solved by current models. In order to proceed, extra information gathered from fricative productions is needed.

Spectral analysis is one of the most important tools available when analysing acoustical speech data, since it provides information pertaining to the source and resonant characteristics — and hence aspects of the shape — of the tract. For vowel analysis, spectral methods have been straightforward to use, and usually provide a clear picture of many aspects of the behaviour of the production mechanisms. However, fricative spectra have a large variance if the signal is not treated properly. This variance can swamp features of interest. The feasibility of using time and ensemble averaging techniques to reduce the variance is examined, but fricative productions can be considered neither stationary nor ergodic, and so these averaging techniques are limited. Frequency smoothed estimates are explored, but found to be of limited use, since they are biased in regions where the spectrum is not flat.

Multitaper analysis is an alternative method of generating spectral estimates with reduced variance, without relying upon assumptions of stationarity or ergodicity, and which provides accurate information pertaining to spectral features. It is optimal over short segments of stochastic data, and so variations in the spectrum over time, as well as over tokens can be estimated.

In order to gain a first estimate of typical variations across productions, to which abnormal productions can be compared, it was necessary to analyse some 'normal' speech. Recordings were made of six normal hearing subjects of each gender, and of Southern English accent, reading a corpus of real words containing /$V_1FV_2$/ segments, where /F/ was one of the eight English fricatives. Six vowel contexts were incorporated, resulting in a set of 3,456 fricative tokens. Of these, the 1,728 voiceless fricative tokens were used in an extensive analysis. In addition, recordings were made of two male and two female postlingually deafened subjects fitted with cochlear implants reading a standard corpus of real words.

Spectral moments have become a popular method for characterising the overall shape of fricative spectra that have a large variance. The parameters with which the moments are calculated are explored, and it is shown that when frequency range, magnitude scale and 'zero reference' are chosen carefully, stable moments that can separate the sibilants can be generated. A high correlation between the odd moments, and the even moments is found, and so the first two moments are best to consider. However, no other improvements can be made, and spectral moments are shown to be insensitive to some changes that are clearly significant when other tools such as spectrograms are used.

Multitaper spectra are used to develop several new parameters that allow for improved classification by place, and characterisation of spectral variance. These analyses provide extensive new information pertaining to fricative production, which is straightforward to interpret. Results of across speaker, within-speaker, within vowel-context, and within-token spectral variations are presented for all the voiceless fricatives. Correlations of overall spectral intensity to spectral shape, and spectral correlations are also shown.

# Acknowledgements

I am indebted to my supervisor Dr. Christine Shadle, who has given her time and knowledge so generously. Also to my family, whose tremendous support and advise has made it all possible.

Thanks also to the Engineering and Physical Sciences Research Council for issuing award number 00310948, to the Electronics and Computer Science Department of the University of Southampton, and to the South of England Cochlear Implant Centre.

Salutations also to friends and subjects M.H., N.S., J.S., W.B., M.G., P.M., S.C., J.B., A.T., Z.S., A.B., N.C., D.B., Carl Verschuur and Julie Brinton, many of whom stepped forward in my greatest hours of need. Huge thanks also to Dr. Adam Prügel-Bennet and Gordon Ramsay for their technical expertise and reassurance, and also to Joanne May, Amanda Goodacre and Abigail Rothwell.

Finally many thanks to my examiners Prof. Martin Cooke and Prof. Mark Lutman, and also to Prof. Bob Damper, for their most valuable feedback and suggestions.

*For Andrew, Rosanne, Mark, Katie and Elina*

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In this work, the variations in normal fricative productions are characterised, with a view to being able to better describe productions made by speakers with impaired hearing.

It is known that hearing perception plays a role in monitoring speech production, although in precisely what capacity is still largely debated. This is easily demonstrated by hearing impaired speakers, whose speech is often disordered. A hearing impaired speaker may have some level of hearing restored with the use of some artificial device (for example, a cochlear implant), and often this will affect the speech, as well as the speech perception, of the subject.

Evaluation of the speech of such speakers is important in terms of both understanding what role hearing plays in production, and also how devices such as cochlear implants may be improved. While the speech perception of such a subject at any stage of hearing ability can be evaluated quite easily, measurement of the quality of the speech production of the subject presents some significant obstacles.

The characteristics of the acoustical speech signal determine what is perceived by a listener. A firm understanding of these characteristics must be obtained in order to better understand normal perception, as well as improve aided perception.

The acoustical speech signal is generated by the vocal tract, and so, at any point during production, the articulatory state of the vocal tract determines the characteristics of the signal being produced. The perception of speech, then, is governed by the acoustical signals that are in turn controlled by the state of the vocal tract.

Different speech sounds are classified by the *manner* in which they are created by the vocal tract. Speech sounds created in a certain manner will have characteristic acoustical features; further, these different sounds are also perceived as distinct. Adjustments within a particular manner of production are known as *place* changes, and will result in changes in the properties of the acoustical signal.

In other words, voicing, airflow, the manner of production, and place of production determine the acoustical waveform produced, and this waveform will determine the perception of a listener.

As a first hypothesis, it should therefore be possible to determine both the manner and place of production, as well as the expected perception of a sound, from the acoustical waveform alone. Indeed, for certain classes of speech sounds, this is true to a certain extent. The appearance of several strong formants in the acoustical waveform allows the approximate positions of the vocal articulators to be determined, and the expected perception of the signal can be hypothesised.

The view above is rather a simplified one however. In fact, for a given phoneme, the articulatory setup can vary to a great degree. Often, the context of the phoneme, the rate at which the talker is speaking, the emphasis of the particular word, the ambient noise and many other factors each contribute to exactly how the phoneme is formed. Each of these articulatory changes will contribute to some change in the resulting acoustical signal, although the perception of the acoustical signal will usually remain constant.

Additionally, certain distinct manners of production, while easily distinguished by listeners, produce waveforms with apparently very similar characteristics. An example of these are the *fricatives*. They are produced when the vocal tract forms a narrow constriction through which air is forced. This generates turbulence noise. The English fricatives are the voiceless /f,θ,s,ʃ/, and their voiced counterparts /v,ð,z,ʒ/. (While /h/ is often described as a glottal fricative, its place varies significantly with phonetic context, making it acoustically variable; perhaps for this reason, it is often excluded from studies of English fricatives.)

A solid understanding of fricative acoustical production has yet to be found. While traditional source-filter models have provided some insights into production, they still leave many questions unanswered.

Moreover, perhaps as a result of their stochastic nature, distinguishing characteristics in the acoustical signal among the fricatives have remained largely allusive. Usually, they are either subject to variations which make them confusable with other fricatives, or are entirely indistinguishable.

New methods for analysing fricative signals need to be explored. A review of much of the work done in these areas follows.

## 1.1 Evaluating normal and disordered fricative production

The speech of subjects with some degree of hearing loss is often examined to help clarify the role of hearing in speech production (Lane, Wozniak, and Perkell 1994). Having a means to measure the improvement or reduction in performance of a subject also allows decisions to be made about the subject's response to various treatment (e.g. Lane and Webster (1991)).

The development of cochlear implants (CIs) has allowed those who are profoundly deaf to have a certain level of hearing-function restored. Initially, limitations in the physical production of these devices determined how much, and the manner in which information could be supplied to the auditory nerve. While improvements in this area continue, it no longer seems to be the limiting factor in restoring hearing-function. It now becomes the task to ascertain a more fundamental understanding of the intricacies of the normal hearing system, if further improvements are to be made.

The perception test scores of cochlear implant patients can return to very high levels following cochlear implantation, but this is not the case for all subjects. Additionally, the speech of these subjects can return to near-normal quality. In some cases however, the speech of a cochlear implant user will continue to be degraded in some aspect.

For a given aspect of disordered speech, it becomes desirable that the main cause for this degradation in performance be found. If the implant is not providing the auditory nerve with a particular important cue, the speech is likely to be affected in the long term. Accurately measuring the way in which the speech production is affected therefore becomes an important task.

Consider the fricative production of a cochlear implant user. The cochlear implant can perhaps be considered as optimised for supplying the auditory nerve with information on formant positions (Wilson 1993). Various 'processing strategies' have been attempted, and while most of these attempt to deliver at least some of the information present in fricatives, it seems possible that this area is the source of some perception difficulties (Matthies, Svirsky, Lane, and Perkell 1994).

In order to evaluate the fricative production, the resulting acoustical signal can be compared to a 'normal' speaker's production. In this case, the aspects of the acoustical signal that reflect the production behind it are required, so that a comparison to normal productions can be made[1].

Measuring normal production variation is therefore an important first step to describing disordered speech. The normal range of productions made by a single speaker, as well as the variations typical across a number of speakers, must be known.

But what aspects of the acoustical signal should be measured? Many attempts have been made to determine which aspects of the acoustical signal represent states of the various production mechanisms used to form it. For vowel sounds, the formant positions of the acoustical spectrum suggest the shapes and sizes of various cavities in the vocal tract. For the fricatives however, the problem appears to be much harder.

The acoustical properties of fricatives are rather different from sounds such as vowels. The acoustical differences result from the fundamental differences in which they are produced. The main source of excitation of the tract during vowel production is a pseudo-periodic signal generated by the glottis. During voiced fricative production, additional sources are set up at forward positions in the tract, and these additional sources are stochastic in nature. In the case of voiceless fricatives, the glottal source is dropped altogether.

If air is forced through a sufficiently constricted passage within the tract, the airflow becomes turbulent, and this generates noise sources in forward positions. These differences in tract excitation sources lead to a number of significant differences in the resulting acoustical waveforms. Due to the nature of the turbulent sources, excitation continues at much higher frequencies than for vowel production. Further, the resonant chambers posterior to the noise source are not excited as much as those anterior, often resulting in much lower energy at lower frequencies (O'Shaughnessy 1987): most energy exists above 2.5 kHz for the palatal fricatives, above 3.2 kHz for the alveolar fricatives, and with very little energy at any region in the spectrum for the labial and dental fricatives.

[1] A method that accomplishes this task well may also be capable of good phonemic discrimination. The task of phonemic discrimination on its own though, while having uses in various automatic speech recognition (ASR) systems, does not have to satisfy the criterion of being able to discriminate the means of production, which is an important task in speech and hearing research.

## 1.2 History

### 1.2.1 Acoustic description and classification

Attempting to locate the important cues in real fricatives is therefore essential to both the study of understanding perception, and also production. It is also the task of work involving other automated systems such as automatic speech recognition (ASR), and voice verification systems.

An important subtle difference should always be borne in mind when discussing cues. *Perceptual* cues are the most likely to yield good ASR performance, and better understanding of the human perceptual system. However, the perceptual cues will not necessarily relate to the means of production. On the other hand, *production* cues will contain information that correlates strongly to the manner and place of production, while not necessarily being significant in terms of their perceptual effects. To highlight this distinction, the different viewpoints of fundamental production-perception theories can be considered, and these are discussed briefly in §1.2.4.

Hughes and Halle (1956) stated that since the voiceless and voiced fricatives /s,ʃ,f,z,ʒ,v/ are easily classifiable by normal hearing subjects in any phonetic context — real or nonsensical — it can be presumed that the perceptual cues for these fricatives lie wholly within the acoustical signal. The spectra of the central 50-ms portions of these fricative were then examined for such cues[2]. Fricative segments were gated from the centre of productions of normal words spoken by three speakers. When the spectra were examined, the distinctiveness of different fricatives could only be described in fairly crude terms: the total amount of energy in certain arbitrary frequency bands, and the magnitude of a frequency peak in a band; no attempts were made to explain the patterns in terms of the production mechanism. A procedure for the automatic classification of central fricative segments was implemented, based upon the observations of spectra for the three subjects. The testing of the automatic classifier was performed on fricative segments acquired in the same manner as before, with the same three speakers, plus two new speakers. High success rates of such testing procedures are to be expected, but are not a particularly useful measure of the goodness of the classifier. In the next stage of experimentation, listening subjects were asked to classify the same fricative segments as were supplied to the automatic classifier; of interest was the finding that where productions were incorrectly identified by the human listeners, these productions were usually also misidentified by the automatic system.

In another early work, Strevens (1960) attempted to explain differences in the voiceless fricative spectra of /ɸ,f,θ,s,ʃ,ç,x,χ,h/ in terms of place of their production within the vocal tract. It was acknowledged that for speakers to produce a given fricative, many different articulatory postures exist, and therefore care should be taken when describing precise manners of production. The production reasons behind the common *intensity* differences across fricatives are considered, and an assumption is made that the spectral shape will not be altered greatly for a given fricative when produced with different levels of air-flow. In contention with previous findings (Harris 1954; Hughes and Halle 1956), it was reported that listeners experienced no difficulty in classifying the voiceless fricatives spoken in isolation. However, it seems that in this perceptual test, listeners may have had access to the start and end transitions. Productions by thirteen subjects of nonsense words with sustained fricatives were used in order to avoid unwanted 'spurious components'

[2] An examination of /θ/ was not performed since Harris (1954) had already demonstrated that the perceptual cues for differentiating the non-sibilants were mostly confined to changes in the formants in the vowel-transitions.

(transitional effects) that would be present in real-word speech. Analysis of amplitude spectra was found to be of little use, since peaks in one spectral slice will often conflict with another spectral slice of the same utterance. However, spectrograms of each production were found to be more useful, and an "average line spectrum" — indicating the main high-energy regions — was found to be the best descriptive measure. Various analyses up to 12 kHz (the limit of recording capabilities at the time) were performed in order to locate the upper frequency limits of fricative production, and evidence was found that information may exist above this point. Similarities in the spectra of fricatives with the same place of articulation were found.

*Template matching* of the spectra at vowel-stop transitions was developed by Blumstein and Stevens (1979) as a classification metric to support speech production and perception theoretical hypotheses. Spectra up to 5 kHz were classified into three broad description templates: The 'diffuse-rising' template (where spectral energy peaks are greater magnitude in the high-frequency region), was found to positively identify alveolar burst spectra with about 76% accuracy. The 'diffuse-falling' template (describing spectra where peaks have greater magnitude in the low-frequency region) was able to positively identify bilabial closure and burst spectra with around 77% accuracy. The 'compact' template (where a single spectral peak dominates in the centre of the spectrum) was found to positively identify around 75% of velar burst spectra. When combined with the correct-rejection scores, the classifier was able to achieve positive classification scores of these three classes of around 85%. The results were used to investigate the effects of vowel-context (coarticulation) on spectra, and to provide evidence for acoustic invariance. Although it was found that the templates were better suited to classification in some vowel contexts than in others, it was argued that the overall success rates suggest an acoustic invariance. However, these conclusions are made without consideration of other theoretical frameworks (see §1.2.4).

In the continuing search for a context-invariant measure of stop place, *locus equations* were a production-motivated approach developed by Sussman, McCaffrey, and Matthews (1991). The technique uses straight line regression fits to data points formed by plotting the vowel's second formant onset frequency against the vowel's second formant target frequency. Locus equations then, do not attempt to describe the noise signal in any way, merely the formant transitions occurring in the adjoining vowel, and so are of limited use in non-vowel contexts. The statistical calculations of goodness were again made using the learning data, and only revealed trends across the stops. In later work (Sussman and Shore 1996), locus equations were tested to see whether they could serve as indicators of place of articulation for obstruents, including /s/ and /z/. Again, trends were observed, but firm evidence of discriminatory ability was not found. More recently, Löfqvist (1999) used a magnetometer system to measure the articulatory movements to investigate the correlation between locus equations and coarticulation between consonant and vowel in CV sequences. This investigation found no evidence to suggest that locus equation slope serves as an index of the degree of coarticulation between consonant and vowel.

The need for ASR machines to be able to discriminate across the fricative using as few parameters as possible was outlined by Jassem (1979). With this aim in mind, a completely different approach was used: methods for classifying a large database of 1035 fricative spectra (/f,s,ʃ,ɕ,x,v,z,ʒ,ʑ/) taken from Polish nonsense words spoken by three male subjects were implemented using a highly heuristic approach, without incorporating any knowledge of production whatsoever. High-order polynomial curves fitted to spectra, together with broad energy-band measures, and a 'centre of gravity' measurement (previously described by Weinstein et al. (1975)), were used for the purposes of description. Again, the performance of the classifier was tested on the learning data set, giving a result of very limited use.

The methodology of using spectral 'centre of gravity' as a spectral characterisation measurement was later expanded to include higher-order moments. Forrest, Weismer, Milenkovic, and Dougall (1988) developed *spectral moments* to characterise voiceless obstruents. The first four moments: spectral mean (centre of gravity, centroid), variance, skewness and kurtosis, of normalised spectral density plots on linear frequency as well as Bark-scale frequency axes up to 10 kHz were considered. However, amplitude data were discarded[3]. The classification abilities of spectral moments were evaluated properly for the voiceless obstruents, using data on which the system had not been trained. Spectral moments were found to be capable of good classification of the voiceless obstruents. However, discrimination of the fricatives /f,θ,s,ʃ/ was not so successful. When testing the discriminatory capabilities of spectral moments, even when using the training set, discrimination was found to be poor. The Bark-scale moments were found to give slightly better results. It is suggested that since fricative intensity is not incorporated into the measurements, it may well provide crucial extra discriminatory data. However, when the sibilants were considered in isolation, the classification performance of spectral moments was good, and when tested on unseen data, showed 95% success rates for this task. It was found that the skewness measure was most responsible for these discriminations. In an attempt to improve the discriminatory capability across /f/ and /θ/, a spectral slice from within the transition region of the fricative was incorporated. While improving the results slightly, they remained very poor.

The limitations of LPC analysis — a strong tool in vowel peak tracking and synthesis — were summarised by Wrench (1995), who developed a *multiple centre of gravity analysis* (MCA) approach to classify fricatives, and compared his results to LPC peak-picking analysis and single 'centre of gravity' measures. However, this approach did not yield particularly useful results.

The problems of the large-variance spectral estimates commonly being used in fricative analyses was investigated by Shadle, Moulinier, Dobelke, and Scully (1992). A corpus incorporating the fricatives /f,v,θ,ð,s,z,ʃ,ʒ/ was generated and spoken by two speakers. Sustained-fricative contexts were used to generate *time-averaged* spectra, and /VFV/ contexts were used to generate *ensemble-averaged* spectra. Some of the issues involved when generating time-averaged and ensemble-averaged spectra were discussed: an assumption of stationarity is required in the case of time-averaging, and of ergodicity in the case of ensemble-averaging. Additionally, averaging techniques require the labelling of 'events' in the time waveform, which can sometimes be problematic. Evidence that the fricatives are nonstationary is presented, and the role of vowel context on this nonstationarity is discussed. The benefits of using reduced-variance spectral estimates are also highlighted, in terms of the increased clarity of formants, and (hence) describing differences between the non-sibilants.

In later work, the robustness of spectral moments was investigated by Shadle and Mair (1996), who noted that moments are sensitive to the frequency range considered, as well as the 'effort' level of the speaker and in some cases, vowel context. A maximum frequency range of 17 kHz was considered in the analysis. It was found that the variations that typically existed across tokens within a fricative were generally greater than across fricatives. The conclusions of this

[3] It was also not clear at which amplitude the 'zero' reference lay. This is discussed in §2.3 and §5.1.3.

study were that spectral moments do not reliably distinguish the fricatives. The interpretation of this study was in contrast to that of the results found in the study of Jongman and Sereno (1995): locus equations and spectral moments were investigated in terms of their classification ability of /θ, ð/ and /f,v/, at vowel onset and offset of the fricative. Using only three subjects, and without testing on unseen data, perhaps unsurprisingly, clear *mean* distinctions across the non-sibilants (in terms of spectral skewness and kurtosis, and locus equation slope and intercept) were found.

In later work, Jongman, Wayland, and Wong (2000) reviewed several of the major descriptive methods, including fricative duration, overall and 'relative' amplitude (the change in amplitude from preceding vowel), spectral moments and locus equations. Spectral estimates are made at the beginning, middle and end of the fricative, as well as at vowel onset. In an attempt to improve the performance of these methodologies, some adjustments are made. The window-length used to calculate the spectrum is increased from 20 to 40 ms; it was argued that the resulting "better resolution in the frequency domain at the expense of resolution in the temporal domain" is preferred due to the "relatively stationary articulatory configuration" during fricative production[4]. Analysis was performed up to 11 kHz, on ten male, and ten female subjects, each speaking the mostly-nonsense words /FVp/ where F=/f,v,θ,ð,s,z,ʃ,ʒ/ and V=/i,e,æ,ɑ,o,u/. The mean values of spectral peak location, spectral moments, locus equations, overall amplitude, relative amplitude and noise duration were presented, and Bonferroni tests were made to test the confidence that these means were generated from different distributions. Bark scale moments were reported to be negligibly different from linear scale moments, and so were not presented. Claims were made about the distinguishing ability of many of the measurements, but the statistical methods used to draw these conclusions may be questioned. Certainly, trends emerged, but no data were presented of variability within the groups. In order to find their total classification ability, discriminant analysis was performed using all measures. The accuracy for non-sibilant identification was reported to be 66%, while for sibilants it was 88%. Further analysis was performed to find which of the measures were contributing most significantly to the classification, and it was found that spectral peak location, normalised amplitude, relative amplitude, and spectral mean at fricative onset and midpoint contributed most significantly to this classification rate. While this study provided useful trend data across the fricatives, the concluding remarks that several of these measures serve to distinguish all places of articulation is inaccurate. The often contradictory conclusions within the literature about the distinguishing capabilities of the various measures is also noted by Ali et al. (2001)[5].

Jesus and Shadle (2002) developed some of the spectral measurements that had been outlined in some earlier studies. Portuguese fricatives /f,v,s,z,ʃ,ʒ/ in sustained and nonsense /VFV/ word contexts were spoken by two male and two female subjects. Spectra were calculated using 10-ms windows, and both ensemble, and time-averaging methods were used to reduce the variance of the spectral estimate. High-frequency and low-frequency '*spectral tilt*' were measures of the mean spectral slope on either side of, and intersecting, the frequency of maximum amplitude. Additionally the '*dynamic amplitude*' (the minimum amplitude below 2 kHz subtracted from the

[4] The claim of stationarity within fricatives had not been properly tested however, and was contrary to the findings of Shadle et al. (1992). Additionally, it is not clear why the 'improved' spectral resolution should improve the performance of the spectral moments.

[5] Unfortunately, the Ali et al. (2001) study goes on to exemplify the seemingly common (yet inaccurate) practice of testing the discriminatory capabilities of one's classifier using the learning data set, and publishing the results as classifying ability (or more impressively, success of determining production place). This topic is discussed further in §1.2.5.

maximum amplitude above 500 Hz) was also measured. Some interesting trends were observed in relation to 'effort level' at which the fricatives were spoken, and although the sibilants could be separated using these measures, the small subject set should be borne in mind. The non-sibilants remained inseparable.

A certain amount is now known about the distinguishing features in the spectral shapes of the fricatives. The sibilants generally have a broad energy peak, in the 2.5–3-kHz region for /ʃ,ʒ/, and in the 4–5-kHz region for /s,z/. The non-sibilants generally have a much flatter spectrum than the sibilants, but no characteristics could be seen to distinguish between /θ,ð/ and /f,v/. While it is suspected that valuable information lies in the transition regions, this has not yet been successfully captured by a reliable measure.

### 1.2.2 The study of fricative perception

Many of these studies use artificial, synthesised speech for investigating fricative perception. Using synthesised speech of course allows the signal under test to be manipulated, and hence completely controlled. By making adjustments to the signal, and observing the perceptual responses produced by subjects, some understanding of perceptual cues can developed.

However, where an artificial fricative is being used, adjusting some parameter of this fricative will inevitably also adjust other cues, which both may be unknown, and important to perception and production. It is also occasionally taken that a specific synthetic parameter represents some 'articulatory measure', and the perceptual effects of adjusting this parameter are often misrepresented. Care must therefore be taken when using synthetic fricatives, since it is impossible to eliminate the possibility that alternative causes lie behind observed effects.

The use of real speech means that no 'artificial' cues will be present. However, as before, when manually manipulating real speech in some way, the same problems are generated: by introducing a notch-filter, or temporal break in the signal, many other cues are additionally being generated, and it should always be considered that these additional unidentifiable changes are the chief cause for any results.

Studies of fricative perception have found a number of clues as to the nature of important cues for discrimination. It has been found that some major perceptual cues, especially for the non-sibilants lie in the transition regions of the fricative, rather than the steady-state portion (Harris 1954). This result has implications for fricative description and classification methods.

Perceptual classification of the fricatives /s,z,f,v,θ,ð/ in 17 preschool children was investigated by Abbs and Minifie (1969). Since it had been recognised by previous studies that cues for some of the fricatives lie in the vowel transition regions, stimuli were in the form of *unedited* /FV/ or /VF/ nonsense words, i.e. a complete, naturally produced signal, by a single male adult. The stimuli were presented in pairs, so that a *confusability* table could be constructed. Highest discrimination error rates existed between /f-θ/, and between /v-ð/, where neither voicing differences nor obvious spectral differences exist. To gain a further understanding of when discrimination errors were occurring, the stimuli were analysed for fricative and vowel durations, peak fricative amplitudes and centre frequency, and energy bandwidths of the fricative. Sibilant spectra were found to have high density, high frequency and short spectra; these observations were used to explain the high discrimination rates between these and the non-sibilants. Differentiation of the voiced-voiceless seemed to be facilitated by the reduced duration of voiced fricatives.

The perceptual effect of differing degrees of variability in /s,ʃ/ productions across twenty speakers in a /CV/ context was investigated by Newman, Clouse, and Burnham (2001). Recordings were filtered to 9.5 kHz, and spectral analysis was performed on a moving 15-ms window, incremented in 5-ms steps, and commencing at frication onset. Mean and variance values for each of the four spectral moments were calculated for each of these windows following the procedure outlined by Forrest et al. (1988). When listening to the *unedited* tokens, it was found that the task of discriminating the sibilants of speakers with more variable productions (in terms of spectral moments) took longer. Conversely, discrimination of speakers with more distinct productions, was quicker. In both cases accuracy remained consistently high. It was concluded that, when present, spectral moment-like cues are used to distinguish the fricatives, but when these are insufficient, other cues, perhaps in the vowel transition region are used.

In some cases, real speech is edited in a limited manner. In a study by Yeni-Komshian (1981), the sibilants /s,z,ʃ,ʒ/ in /FVFVFV/ contexts (where F=/a,i,u/), spoken by one male and one female, were edited so that fricative, and vowel-transition regions were isolated. Various portions of the vowel and/or fricative were played to eight subjects, in an attempt to learn the effects of coarticulation on the perception of both. Strong evidence was found that fricative-vowel coarticulation affects the perception of the sibilants[6]. The conclusions of this investigation were tested by Jongman (1989) who considered a larger set of fricatives (/f,s,θ,ʃ,v,z,ð/) in a similar manner. Fourteen subjects listened to extracted portions of the productions of a single speaker. Plots of 'relative information transmitted' were presented to show how the increase in (duration of) information to the listening subjects related to percentage chance of successful classification. It was also noted that identification of place of articulation was much more affected by fricative duration than were manner and voicing. Another similar investigation of the location of perceptual cues for discriminating /f/ and /θ/ which remain largely unclassifiable was later performed by Hata, Moran, and Pearson (1994). This was done using perceptual experiments of isolated segments of the fricatives F=/f,θ/ (excluding any vowel waveforms) in /FVF/ contexts produced by a female speaker, and downsampled[7] to just 10 kHz. It was found that when presenting the frication portion alone, /f/ was identified correctly more often than /θ/. When the entire following vowel was included, identification of /f/ significantly improved, while identification scores for /θ/ were unchanged. It was found that more than 30 ms of the following vowel needed to be included in order for the perception to improve. Perception scores remained somewhat low even when the entire vowel was presented, and it was suggested that the frequency range being used may have been insufficient for discrimination. It must be considered that the effects of manipulating speech in the artificial manner in these studies may have had a large uninterpretable effect on their results and hence findings.

Although speech perception scores remain high even when rather drastic lowpass filtering is implemented, it is entirely possible that higher-frequency cues exist and contribute to a greater degree when lower-frequency information is missing. The suggestion of fricative perceptual cues lying in higher-frequency ranges may also contribute to the poor performance of perceptual

[6] Also included is a detailed discussion about the possible causes of classification failure regarding those portions of each production that were removed, which are consistent with the theories of Kent and Minifie (1977).

[7] Presumably after bandpass filtering to 5 kHz.

studies of fricatives under lowpass conditions. The role of these higher-frequency cues was investigated by Lippmann (1996). Nonsense /ə-CVC/ syllables (where C ∈ /p,t,k,b,d,g,f,θ,s,ʃ,v,ð,z,ʒ/ and V ∈ /i,ɑ,u,ɪ,ɛ,ʊ/) were spoken by an adult female. Six notch filters were used to manipulate the recorded tokens. Each notch filter had a passband from DC to 800 Hz. Upper passbands were from $f_u = 3.15$, 4, 5, 6.3, 8 and 10 kHz up to 20 kHz. When subjects were asked to identify the consonants under different notch-filter conditions, it was found that performance fell smoothly from 91.6% for $f_u = 3.15$ kHz to 73.9% for $f_u = 8$ kHz. It then dropped more sharply when more high-frequency information was discarded. When analysing the fricative perception alone, a number of interesting observations were made. Perception of /s/ was near 100% for values of $f_u \leq 8$ kHz, after this, scores fell sharply. Scores for /z/ were also near 100% up to $f_u = 6.3$ kHz, after which they fell sharply. However, for the fricatives /ʃ,ʒ,θ/, scores were only high for values of $f_u \leq 4$ kHz, after which they fell steadily. These findings agreed with those previously discovered by Lacerda (1982), that eliminating the apparently most distinctive spectral features had little effect on the perception of the fricative. This has important implications for the way we think about speech perception, and hence, the assumptions about the nature of 'cues' in speech, although the artificial editing of the speech sound must be considered as having an effect itself.

Fricative synthesisers have important roles in both commercial and research applications. The only manner in which the 'successfulness' of such a synthesiser can be measured, is by testing human perception of the resulting sounds that can be produced. In their development of such a synthesiser, Heinz and Stevens (1961) used a model of the vocal tract, consisting of a single noise source, and a pole-zero filter arrangement. When isolated stimuli consisting of single bands of noise with varying centre frequency was presented to subjects, differing responses were elicited. Centre-frequencies around 2–2.5 kHz produced /ʃ/ responses, around 5 kHz produced /s/, and around 8 kHz produced both /f/ and /θ/ responses. Synthetic fricative-vowel syllables were then generated, and agreement with the findings of Harris (1954) was established concerning the importance of vowel transitions in the perception of non-sibilants. These findings were in agreement with previous studies (Harris 1954; Hughes and Halle 1956).

Many perceptual experiments make use of synthetic fricatives in order to try and establish perceptual cues (e.g. Gurlekian 1981; Stevens et al. 1992; Hedrick and Ohde 1993; Cheesman and Greenwood 1995; Formby and Childers 1996). Other studies use both synthesised and natural speech stimuli, occasionally concatenating the two (e.g. Nittrouer and Studdert-Kennedy 1987; Zeng and Turner 1990; Whalen 1991; Johnson 1991; Nittrouer 2002). As previously discussed, great care must be taken over the interpretation of perceptual results using artificial stimuli. In some cases, the use of highly-controlled synthetic fricatives can be justified, but their use is often exploited beyond interpretable limits. The findings of many of these studies could easily be interpreted differently to the conclusions published. One of the most commonly-used synthetic-speech analysis tools is the *continuum*: a single parameter of the stimulus waveform can be varied through a range of values. When this parameter is adjusted, the perception of the sound invariably also changes; but from this care should be taken over drawing conclusions that the specific artificial 'cue' that was adjusted was solely responsible. That is, the adjustment of one artificial parameter will invariably alter many other possible cues. In general though, these studies provide useful supporting evidence of the existence of perceptual cues in the vowel-transition regions, and also of the effects on perception of the interdependency between cues (such as relative amplitude and spectral shape (Hedrick and Ohde 1993)).

Evidence of the problems faced when using edited natural, or synthetic speech can be seen by designing a perceptual experiment in which both sources are used as stimuli. The disparate results of such a study were briefly mentioned by Stevens et al. (1992), although this was considered insignificant.

Some studies attempt to model the important acoustical properties of fricatives, and observe the effects on perception of changes in 'acoustical cues'. However, since these changes do not generally relate to changes in tract configuration, the effects on perception are of limited value.

A more feasible approach would appear to be that of modelling production and observing the effects on perception of 'changes in production'. The synthesisers used in such studies generally rely upon source-filter models of the tract that work well for vowel synthesis. However, these models do not generally take into consideration the complex effects of multiple forward sources for given configurations, and therefore produce acoustical signals that may not be representative of the configuration in question. Measures of changes in perception to some artificial change in 'configuration' must also be questioned.

As an alternative approach to tackle the problems of using either synthetic, or artificially edited speech, while maintaining good control over certain 'parameters', Fletcher and Newman (1991) implemented a visual articulatory feedback mechanism. A *palatometer* was fitted to two adult males, allowing observation of the positions of contact with the tongue. These same two male subjects were used in a further perceptual study, in which correlations between contact-positions with the palatometer, and perceived sibilant were calculated. The effects of such invasive measurements of production place are unfortunate, but unavoidable, and nevertheless provide reasonable approximations to normal speech. However, the choice of subjects for the perceptual part of the study meant that the findings of perception of specific place changes are of questionable statistical validity. If the perceptual part of the experiment could be repeated under stricter conditions, highly valuable information pertaining to theories of perception of place could be ascertained.

In a very different approach, the importance of certain acoustical characteristics on fricative perception was examined by testing the speech perception scores of CI subjects with a strategy that emphasised certain features. The '*transient emphasis spectral maxima*' (TESM) strategy developed by Vandali (2001), boosted electrode-potential at times and frequencies of brief temporal features. This emphasis on brief temporal — particularly low-level — acoustical features resulted in an improved performance in fricative perception.

### 1.2.3 Analysis of speech production

This section presents studies that have concentrated on observing and describing the actual production characteristics, rather than the acoustical characteristics of fricatives. Often, metrics discussed in the previous section are used. As mentioned earlier, one of the main purposes of developing tools to measure the 'characteristics' of fricatives is so that variations in production under different conditions can be measured.

The idea that positional information in the vocal tract could be extracted from the acoustical signal was demonstrated by Strevens (1960), who considered that since articulatory configuration was the chief cause of acoustical changes, the measurement of these acoustical signals should — when combined with other physical measurements — contain information about the state of the vocal tract. Crude measurements of speech intensity and of pulmonic air-pressure revealed that the amount of pulmonic air-pressure required to elicit a certain sound intensity varied according to place of constriction. Analysis of spectra revealed that the place of articulation produced distinctive features in the spectrum. *Front* fricatives (labial and dental /ϕ,f,θ/) tended to have the broadest (i.e. least 'peaked') spectra, and generally the lowest intensity. The *mid* fricatives (alveolar and palatal /s,ʃ,ç/) had a more 'peaked' spectrum, with a peak occurring around 3–4 kHz, and generally eliciting the greatest intensity. And the *back* fricatives (velar, uvular and pharyngal /x,χ,h/), had intermediate intensity, and a more 'formant-like' spectral structure.

Soli (1981) studied the spectra of the sibilants /s,z,ʃ,ʒ/ in /F/ and /FV/ contexts (where V=/a,i,u/) in a single male talker. Mean LPC spectra were used to gain a reduced-variance spectral estimate. From the [s,z] productions, peaks in the F2 region of the following vowel were reliably seen before vowel onset, and it was suggested that this may occur when the constriction is insufficient to cancel these resonances. For [ʃ], spectral peaks relating to both F2 and F3 of the following vowel were observed, where F3 generally defined the main peak of the fricative spectrum. It was suggested that the peaks observed in the fricative LPC spectra resulted from back cavity resonances, although the effects of back cavity resonances had previously been thought to be different in fricatives and vowels, due to the antiresonance set up by the forward noise source.

The relationship between the duration of a fricative, and its voicing was explored in the fricatives /f,θ,s,v,ð,z/ of just three male subjects by Baum and Blumstein (1987). Contrary to their expectations, it was found that, while mean durational differences did exist, the variations among them produced significant overlaps, so that very little information concerning the nature of voicing in fricatives could be gained from this measure alone. These findings were confirmed by Crystal and House (1988), who additionally found that fricative duration was also noticeably affected by its position within a word, as well as whether it appeared in connected speech or citation form.

Useful information pertaining to production theories has also been gathered by comparing differences between child and adult speech. Spectral mean (up to 9.6 kHz), and amplitude in the sibilant production of eight children and four adults in /FiFi/ and /FuFu/ contexts was compared by McGowan and Nittrouer (1988). Peaks in fricative spectra thought to pertain to F2 in the vowel spectra were selected 'by eye', and were found to be higher in females than males, and higher in children than in adults. This is explained by the reduced size of the back cavity, and hence increased resonance. Additionally, it was found that these fricative spectral peaks were more more distinctly affected by vowel context in children than in adults. In later work (Nittrouer, Studdert-Kennedy, and McGowan 1989) evidence was found that mean sibilant centroid measurements became more distinct with age, although higher-order moments provided no additional evidence of this (Nittrouer 1995).

Frontally misarticulated [s] productions are often considered to resemble normally articulated [θ] productions, both perceptually, and in terms of articulatory configuration. Baum and McNutt (1990) tested 10 children with disordered [s] production against 10 children with normal [s] productions. Interestingly it was found that [θ] productions in the disordered subjects had mean durations closer to that of normal [s] productions, whereas the durations of [s] productions were similar across both groups. However, the amplitude was found to be a more distinctive measure difference between /θ/ and /s/ tokens in both groups of subjects, yet no significant changes were found in the amplitude data of the misarticulating subjects. In terms of spectral mean, *in both groups* some subjects maintained spectral differences in their [s] and [θ] productions, whereas others did not, and while these distinctions may not be readily heard by a listener, they nevertheless indicate that an internal differentiation between the two fricatives occurs. These findings were a strong demonstration of the power of acoustical measurement tools over perceptual acoustic measurements, and hence, for example, the unsuitability of labelling misarticulated [s] as [θ].

In the study of Fletcher and Newman (1991) mentioned earlier, a palatometer allowed accurate measurements to be taken of the place of constriction in /s,ʃ/ productions by two male speakers This positional information was compared across the subjects, and it was found that the sibilants were produced in quite different positions on the alveolar ridge across speakers, although groove width was more consistent.

A closer analysis of the vowel transitions and voicing differences in the fricatives /f,v,s,z/ in different contexts, from acoustic analysis (up to 4.8 kHz) was performed by Stevens, Blumstein, Glicksman, Burton, and Kurowski (1992). Measurements of duration were in accordance with previous studies: voiced fricatives on average being around 30 ms shorter than unvoiced (although this was a smaller difference than found in previous studies), and the preceding vowel being longer when followed by a voiced fricative. The duration of a fricative is longer when in utterance-final context (around 41 ms) and shorter when in inter-utterance context (around 24 ms). Progressive 30-ms Hanning windows were used with a 20-ms overlap, and running approximately from mid-vowel to mid-fricative positions. As expected, the F1 peak reduces in amplitude over /VF/ transitions, and increases over /FV/. This F1 peak — thought to represent glottal vibration — sometimes continued after frication had commenced in unvoiced fricatives, but also often discontinued in the central regions of voiced fricatives. This led to conclusions that features other than the duration of glottal vibration must be present in signaling the voicing feature in fricatives.

Spectral moments were used to measure the degree of production variability in the sibilants, within and across subjects, by Newman, Clouse, and Burnham (2001), as mentioned earlier (§1.2.2). Of the four moments calculated, the centroid (M1) and skewness (M2) measures were found to differentiate the fricatives to the greatest degree. The study clearly demonstrated that within a speaker, the variations of M1 and M3 for a given sibilant would usually not overlap the variation of values for the other sibilant, although this was not always true. In isolation then, these measures were seen to be insufficient at completely distinguishing productions of different sibilants, suggesting the existence of alternative cues. However, they also highlighted the significant degree of variations in production within a speaker, and within a given fricative. Although the sibilants were usually clearly separated by these measures within a single talker, when comparing across speaker, it was found that in some cases, the values for one subject's production of [ʃ] would completely overlap with another subject's [s] values. When combined with the previous suggestion, it seems most probable that other cues must exist that are not being considered. However, in the perceptual portion of this study, listeners took longer in classifying the sibilants of speakers with less distinct (i.e. containing large overlap) spectral moments. This could suggest that, while serving as the primary cue, alternative cues are used if this cue is insufficient. In a further analysis, it was found that a high statistical correlation existed between the centroid and skewness values of a given subject, although this was considered as a reinforcing role[8]. Further investigation of the within-speaker variations in sibilant productions was undertaken by Munson (2001). The centroid was considered over time in the sibilants in limited contexts. Evidence was presented that sibilant production variability depends upon context.

When the hearing system is altered by deafness, and following artificial stimulation, analysis of speech production may lead to clues about the hearing system's role in production. Lane and Webster (1991) demonstrated the role of self-hearing in speech production by looking at the deterioration of postlingually deafened adults. A form of spectral mean (Jassem 1979) was used for the purposes of measuring differences in [s] and [š]. Measurements of the centroid showed a reduced production distinction between the sibilants compared to normal hearing subjects. When limited hearing capability is restored by artificial means such as cochlear implants, measures of speech production often move towards 'normal' values (Lane, Wozniak, and Perkell 1994). Matthies, Svirsky, Lane, and Perkell (1994) used spectral moments to demonstrate the improvements over time (i.e. movement towards 'normal' values) in the productions of sibilants in three out of five CI subjects. In a later study (Matthies, Svirsky, Perkell, and Lane 1996), evidence was also presented that the improvements in spectral moment values (centroid and skewness) were probably connected to improvements in articulatory configuration, using a electromagnetic midsagittal articulometer (EMMA).

Clearly, there is much that remains to be discovered about the effects of cochlear implants on speech production. It is important that analysis in this area is undertaken with the utmost care, and with careful consideration of underlying assumptions.

### 1.2.4 Speech production-perception theory

Within the studies reviewed, apparently significant features have been observed at different points along the chain of human communication. Articulatory features, such as manner and place correspond to characteristic features in the resulting acoustical signal. Certain features in the acoustical signal, such as spectral shape, have been found to be important in determining what will tend to be perceived by a listener. And specific changes in hearing ability correspond to different changes in both perception, and production.

However, it must always be considered that the different features that are observed in experimental analysis may, or may not be important to the underlying processes of speech production and perception. Further, even if it can be established that a particular feature is important to these fundamental processes, it is important to consider precisely what role the feature plays. An exciting area of speech research is that of theorising about and establishing the underlying processes of speech production and perception. A number of theories about speech perception and production exist, and it is appropriate to mention a few of the important ones here, since they will inevitably contribute to our interpretation of experimental results.

Although it was originally theorised that speech could be broken down into small phonemic segments that conveyed clear and complete sequences representing tokens, this was quickly found

[8] The calculation of spectral moments may incorporate methodological instabilities that contribute to the degree of variation and inter-correlation. These are investigated in §2.3 and in Chapter 5.

to be a problematic approach to speech perception. The main confounding factor was found to result from phonemic segments overlapping with each other, sometimes to such an extent that a short phoneme could disappear altogether. These overlaps were found to result from articulatory configurations of phonemes overlapping to various degrees. Theories that tried to account for this coarticulation effect, while maintaining the original ideas of ordered sequencing were reviewed by Kent and Minifie (1977). Coarticulation itself was found to be subject to unaccountable variations, and new theories of speech production and perception evolved.

A good comparison of these is presented by Hawkins (1999b). These theories are often quite diverse, and unfortunately often lead to conclusions that cannot be easily tested by experimentation, if at all. They are all, of course, based on experimental results, and modelled on certain findings. However, often different experimental findings are considered more significant, and are explained more convincingly by an alternative theory. Common to all theories, however, is that they must explain *all* experimental results, and though this is usually attempted, explanations are sometimes of questionable plausibility.

For example, the *motor theory*, which has long been one of the more popular suggested theories of perception, relies heavily on some awkward underlying assumptions (Hawkins 1999a). One of the more intriguing problems regarding speech perception is that, in order to transmit a single specific abstract 'token', speakers can and do produce an immense variety of different acoustical signals to represent this 'token', and yet these signals will all invariably be perceived by a listener as the single correct 'token'. How can this vast range of acoustical signals all be interpreted as the single intended token? *Motor theory* suggests that, although the acoustical signal produced to represent a specific token is highly variable, the underlying production mechanism required to do so is not. It is suggested that from the acoustical signal, a special 'speech module' within the brain allows the underlying articulatory movements that were used to generate the signal, to be abstractly 'viewed'. From this invariant articulation abstraction, the intended token can be recognised. However, little is offered in the way of explanation of how the articulatory movements are decoded from the acoustical signal, other than that it is innate to the speech module, and present from birth.

An explanation for the way in which the articulatory configuration can be recovered from the acoustical signal is proposed by Stevens (1989) in his '*quantal*' theory. It is proposed that the acoustical speech signal can be broken down into invariant properties and variant properties. The invariant properties are those relevant to the articulatory configuration, while the variant properties do not contribute, and are 'discarded'. These invariant cues within the speech are enhanced by optimised production and perceptual systems. Specifically, a given important acoustic feature will change nonlinearly with respect to the movement of the articulator, so that the acoustical signal will only change significantly when the articulator also does. A similar arrangement is proposed to exist in the perceptual system. Quantal theory offers an attractive explanation for many findings of production and perception. However, in many cases it does not.

The occurrence of a highly-variable acoustical signal representing a single token was considered from a very different perspective by Lindblom (1983). Rather than viewing the acoustical signal as having invariant 'cues' and variant 'noise', it was considered that in fact the variations that are produced for a single intended 'token' result from additional influences that play a critical role in transmitting the meaning in an optimal manner. A trade-off is set up between producing transmitting clear, easily identifiable acoustical information, and conveying this information in a manner appropriate to the situation. Coarticulation, while reducing the clarity of the intended token, allows the speed of information transmission to increase, ultimately increasing the overall transmitted information. It is suggested that acoustical data are optimised to supplement (and hence clarify from ambiguity) an expected input. *Hyper- and Hypo-articulation* (H & H) theory offers a theoretical production framework that expects the differing degrees of variability that is observed in speech. Investigating these variabilities (rather than the '*invariabilities*') may result in an increased understanding of human speech production and perception.

### 1.2.5 Summary

Perceptual cues that can discriminate the fricatives must exist somewhere within the acoustic signal. However, determining precisely which features of the acoustical signal represent these cues has been a highly problematic area. This is due in part to the large variabilities across productions, combined with the effects of interactions with context of the fricatives. Theories of underlying speech production and perception processes have helped to predict and eliminate some of these problems (e.g. coarticulation), but classical theories do not account for all that is seen.

Some studies consider the production mechanisms behind the signals in order to evaluate them, and attempt to explain spectral shapes in these ways. Other studies take a more 'observational' approach, and use statistical results to locate likely discriminatory cues. The difference between *production* and *perception* cues should be borne in mind for each of these studies.

In some cases, measurements of production are used to establish articulatory information, but how reliable are these measurements at predicting articulatory configurations? Other studies attempt to use various measures of the acoustical signal to speculate about likely explanations for perception capability, but is this really reliable when so little is still understood about the fricatives?

In the study of fricative production, subjects may either be asked to read real word lists, or 'nonsense' words, in some cases producing sustained fricatives, that rarely occur in speech. The use of sustained fricatives is likely to lead to better understanding of the noise portion of the fricative, but is unlikely to produce results related to variations that typically occur in speech. The behaviour of voicing in fricatives is often examined in real or nonsense word corpora, but again, the variations of natural speech are unlikely to occur.

Many of the more basic measures that can be used in speech production usually produce unreliable information. Simple duration and amplitude measurements do not tend to yield informative results, although they often suggest trends.

Traditional spectral methods that have been invaluable in vowel-speech analysis, such as LPC, also become unreliable when considering fricatives. Other spectral methods, such as template matching, have been developed to try and find distinguishing features in the spectrum, but these have been very limited in terms of information they return about production, and classification ability is also limited.

The large variations in fricative spectral estimates have often been seen as both unimportant and problematic, and so broad spectral energy measures have often been favoured. Of these broad descriptors, the most successful have been moments, which capture the overall spectral shape, with little importance attached to narrow frequency peaks. However, a number of problematic issues surround the use of spectral moments. Their broad descriptive capability means that they are unable to capture apparently important finer frequency spectral details, correlations have been found in M1 and M3, and although able to discriminate many sibilant spectra, they often fail at this task, and invariably fail at distinguishing the non-sibilants. Whether these problems result from fundamental limitations in the technique, or from errors in the methodologies is in important issue. For example, some early studies assumed that sound pressure level (SPL) variations would not drastically alter the spectral shape, but more recently this has been shown to be incorrect. There is strong evidence to suggest the interaction between relative amplitude and spectral shape plays an important role in perception. So adjustments to moments may be needed.

It has been found that perceptual cues for discriminating non-sibilants lie mostly in the vowel transition regions. This has been confirmed by automatic speech classification investigations, and also perceptual studies. Occasionally, the transition regions are incorporated into classification methodologies. Of these, locus equations have not been found reliable, and although spectral moments have been used in the voicing region of fricatives, it is not clear whether this usage is justifiable for portions of the signal that contain significant spectral peaks.

However, spectral moments have been used to describe the speech of subjects with some form of speech disorder, or speech affected by hearing ability. In this regard, spectral moments have generally been able to indicate improvements over time in most subjects.

Preliminary studies suggest that valuable information lies in the peak positions and magnitudes of fricative power spectra, but the inaccuracies of spectral estimation have meant that these have been difficult to uncover.

Finally, in the testing stage of many descriptive measures, it has become common practice to test the classification ability on the same set of data as was used to determine differences across tokens. This of course can provide insight into possible significant measures, but is far from establishes what the significant features of all tokens will be. In order to do this, the measures being tested must be tried upon unseen data.

A few additional points remain concerning common practices in fricative analysis. The range of frequencies to be considered in analysis has often been under 11 kHz, and sometimes as low as ~4.5 kHz. There is strong evidence to suggest that fricatives contain significant information above these frequencies. Additionally, practices in obtaining spectral estimates must be reexamined: it now seems common to use estimates with variances that, as will be shown in due course, are extremely significant. While averaging techniques exists, and have been employed in a few studies, alternative methods for obtaining more accurate spectral estimates should be investigated.

## 1.3 Approach

Parameters have been found that partially distinguish some of the English fricatives, but these are commonly subject to large variations across productions. These variations are often so large that the boundaries for different fricatives often overlap, and yet these fricatives remain correctly perceived. This suggests that some of the most important distinguishing cues have yet to be found.

In §2.1 we investigate some of the theory behind fricative production, and the problems facing parametric analysis. Many of the popular fricative characterisation methods involve grouping the energy in the spectrum into broad bands, or describing the overall distribution of spectral energy in very broad terms (such as moments) before further classification stages occur. This approach is well-suited to dealing with the problematic large variances seen in fricative spectra.

The large variances in fricative spectra usually result from first estimates of the stochastic signal that are not consistent. Techniques that exist for reducing the spectral estimate variance are commonly not implemented. Several methods of reducing the variance of spectral estimates in typical fricative signals are investigated, and the practicality of each is discussed in §2.2. Generally, the classical techniques tend to rely on assumptions that do not always hold well. Modern techniques are also investigated, and found to be well-suited to the analysis of fricative signals.

It may be expected that improvements in spectral estimation should lead to an improvement in performance of the more popular fricative classification techniques, such as moments. In §2.3, the fundamental properties of spectral moments are carefully examined.

In order to continue investigating fricative production, it is necessary to acquire suitable test data. The details of the procedures followed are given in Chapter 3. A real-word corpus (given in Appendix A) was devised, and read by six normal-hearing speakers of each gender.

Improved spectral estimates allow more careful observations of typical productions to be measured, and this is investigated in Chapter 4; using these spectral estimation methods, improved spectrograms of fricatives can be generated, and these are presented in Appendix B for the voiceless fricatives in two vowel contexts, for all speakers. The effects of incorporating better spectral estimates into the spectral moment methodology are explored in Chapter 5. Other important aspects of spectral moment calculation do not seem to have been considered in the literature, and these are also explored.

With the variance of the estimate reduced, attempts are made to measure the variation in fricative production in different contexts in Chapter 6. Differences and similarities between fricative spectra are quickly located. Differences across speakers, across vowel contexts, and within fricative tokens are explored, and many of the results of these analyses are given in Appendices C and F. Patterns in production were observed using spectral correlation, which was previously impossible using large-variance spectral estimates; results for male fricatives only are given in Appendices D and E.

The foremost task is not to classify, but to determine which features of the acoustical signal reflect aspects of production. Moments have been useful in the analysis of cochlear implant user's fricative productions where incorrect place is observed. Other problems with production may also exist. Finding methods that enable more subtle spectral details to be described may be useful in such analyses. With measurements of productions from normal-hearing speakers in hand, the analysis of four cochlear implant users is undertaken in Chapter 7.

This work presents significant improvements and new analysis methods that can be made use of in future studies of normal and disordered fricative production, and Chapter 8 concludes with a discussion of a number of possible future applications.

# Chapter 2

# Theory

## 2.1 Fricatives as stochastic processes

In order to analyse fricatives, consideration must be taken over how they are produced, and hence, what characteristics they may be expected to exhibit. Fricatives are produced in a wholly different manner to vowels, and so it is appropriate that a different set of analysis tools may be used. By considering what is known and what is not known about fricative production, tools can be developed that are best suited to analysis of these signals.

### 2.1.1 Turbulent jets as acoustic source

The source of noise is generated near a constriction in the vocal tract. When the air flow is channelled through a constriction, the air particles accelerate, creating a jet of air which has very different characteristics to laminar airflow. Such jets of air have distinguishing characteristics, among which are highly randomised subsidiary vortices and turbulent eddies. These eddies can occur at different places along the constriction and at the exit of the constriction, depending on the airflow, configuration of the constriction and surface conditions. These turbulent eddies generate a random sound pressure source (Meyer-Eppler 1953; Fant 1970; Flanagan 1972). Additionally, the jet of air released from a constriction may be targeted towards an obstacle (such as the teeth) or a surface (such as at the glottis), in which case additional, and often more intense sources of turbulence noise are often produced (Stevens 1998). Rapid transient changes in flow (as for the affricates) can also act as sources of noise (Scully, Castelli, Brearley, and Shirt 1992). We take a moment to review some of the theoretical and analytical fundamentals of turbulence noise sources.

The airflow during fricative production can be treated as an incompressible fluid. This approximation holds well as long as the velocity of the air particles does not approach the speed of sound (Schlichting 1960). During *laminar* flow, particles follow the direction of the tube, or constriction they are within. If the tube is long and straight, the flow will have greatest velocity in the centre of the tube, while at the edges it will approach zero. Fluid particles in the flow are acted upon by the pressure gradient within the tube. Their inertial force is partially determined by the density of the fluid $\varrho$ and the free-stream velocity $V$. Particles also interact with each other due to frictional forces, and these are partially determined by the coefficient of viscosity $\mu$. Both of these forces are influenced by the particle velocity. The ratio of inertial force to friction force hence describes the nature of the flow, and is termed the *Reynolds* number, defined

$$\mathsf{R} = \frac{\varrho V d}{\mu} = \frac{V d}{\nu} \tag{2.1}$$

where $d$ is the *characteristic dimension* (or *effective width* (Meyer-Eppler 1953)) of the tube, and the ratio $\nu = \mu/\varrho$ is known as the *kinematic viscosity.*

The *characteristic dimension* of some tube of arbitrary cross-sectional shape is proportional to the ratio of the cross-sectional area to perimeter. Consider the airflow through some tube: if the volume flow within the tube remains constant, an increase in the *surface area* within the tube (and therefore an increase in the characteristic dimension), will increase the 'inclination' of the airflow to become turbulent. The Reynolds number then, is essentially an index that corresponds to a particular configuration of flow. For an excellent illustrative account of such flows, see Van Dyke (1982).

Because particle velocity varies as a function of distance from the edge of the tube, these frictional forces result in shear forces upon the fluid particles. As the Reynolds number of a system increases, the nature of the flow changes from a laminar flow to a a more chaotic *turbulent* flow. The onset of significant turbulence occurs once a critical threshold $\mathsf{R} > \mathsf{R_c}$ is overcome, where the *critical Reynolds number* $\mathsf{R_c}$ is determined by factors such as the configuration of the constriction and surface properties, and is generally found by experimental measurement. The onset of turbulent flow coincides with the generation of acoustical noise. The mean cross-sectional velocity $V$ of flow within the tube then becomes related to the pressure drop across the constriction $p_d$ as

$$p_d = \frac{\varrho V^2}{2}, \tag{2.2}$$

(Flanagan 1972) (sometimes called the *overpressure*). The pressure drop across the constriction is therefore proportional to the squared particle velocity.

From theoretical and practical analysis, a number of different relationships of far-field sound pressure to overpressure have been suggested. From physical models of fricative production, Meyer-Eppler (1953) demonstrated that an approximation of the sound pressure $p_s$ at a fixed distance from the source of turbulence was given by $\tilde{p}_s \propto \mathsf{R}^2 - \mathsf{R}_c^2$, where $\mathsf{R_c} \approx 1800$ for plastic tube models. From this, the relationship of sound pressure to overpressure was approximately $\tilde{p}_s = k_1 d^2 p_d - k_2$ where $k_1$ and $k_2$ are constants. However, analysis of these results by Stevens (1971) led to the conclusion that the sound pressure was less sensitive to constriction area, and more to overpressure, resulting in the relationship $\tilde{p}_s \propto p_d^{1.5} d$, and this relationship has been used since (e.g. Scully and Allwood 1985).

The shape of a constriction therefore plays some role in determining the resulting sound intensity, and the rate of increase in intensity with pressure. The situation is further complicated by the introduction of some obstacle, or surface incident to the turbulent flow, as often occurs during fricative production, and so more complex methods are required to analyse these.

Monopoles, dipoles and quadrupoles are theoretical representations used to describe sources of noise that are generated by different mechanisms (Landahl 1975), and each of these different source types exhibits different known characteristics (Goldstein 1976). For instance, under certain assumptions, the sound power generated by a quadrupole source is usually proportional to $V^8$, and for a dipole is usually proportional to $V^6$. A turbulent free jet is considered to generate quadrupole sources. When such a jet is directed towards an object or surface, additional dipole sources are constructed. Since each type of source exhibits different characteristics, they are a useful tool when considering fricative production, and are often used to help explain observed behaviour of various models (e.g. Stevens 1971; Shadle 1990; Stevens 1998).

For example, since for low flow velocities (specifically, for $V < c$ where $c$ is the speed of sound), the conversion of kinetic energy of the turbulent flow into sound power is more efficient for dipoles than for quadrupoles, it may be expected that a jet directed towards an obstacle or surface will exhibit greater acoustical intensity. Evidence of such behaviour has been supported by models of the constricted vocal tract by Shadle (1990); the far-field sound intensity was increased by up to 30dB when an obstacle emulating the teeth was introduced into the flow of the air-jet.

A very limited amount is known of the spectral properties of turbulent noise sources. Goldstein (1976) demonstrated that in general, for a free jet, the quadrupole noise source spectrum will be in the form of a very broad peak, the maximum of which is located at a frequency proportional to $V/d$. The most complete analysis of source characteristics in more vocal-tract-like configurations has been undertaken by Shadle (1985), who showed that this overall shape was altered by the introduction of an obstacle (and hence, dipole sources). Whereas for a quadrupole the source spectrum rolls off either side of the maximum, for the dipole-quadrupole combination source, the spectrum generally did not roll off significantly at low frequencies. Additionally, the quadrupole-type source retains its overall shape with changing intensity. Quadrupole-dipole sources are relatively insensitive to changes in flow at low frequencies, but the increase in spectral amplitude grows with increasing frequency.

### 2.1.2 Interaction of source and tract

Unlike vowel production, where the main excitation source occurs at the glottis, which can be considered as one end of the cavity system, turbulent noise set up during fricative production interacts with cavities both posterior and anterior to the sources. These cavities further shape the spectrum in complex ways.

For a fixed configuration, the vocal tract will exhibit specific resonance characteristics. The positions and magnitudes of resonant frequencies (or 'formants') are characteristic of the system configuration. The characteristic spectrum is defined as the response to an excitation occurring at the glottal end of the tract. However, as the source moves to a position significantly forward in the tract, a number of significant changes occur.

Perhaps most significant is that where only resonances existed in the glottis-excited vowel system, when the source is brought forward, anti-resonances may be excited. These anti-resonances are frequencies of infinite impedance looking towards the glottis from the source. These anti-resonances have distinctive effects on the shapes of spectra resulting from forward sources. Specifically, the location of the noise source within the tract has a significant effect on the frequencies of zeros in the system (Fant 1970; Stevens 1971; Flanagan 1972).

A common early approach to modelling the vocal tract with a forward source was the multi-tube approximation (e.g. Fant 1970; Flanagan 1972), based on a series of interconnected tubes. The tubes are intended to emulate the chambers within the tract, and these exhibit resonances and anti-resonances. Electrical circuit representations of multi-tube approximations of the tract are in abundance in the literature, and are good at describing the roles of cavities as resonators, the changes in transfer-function when the tract changes shape, the effects of varying sources, and so on. The models are generally considered accurate up to approximately 4 kHz, beyond which the assumption of plain waves becomes less accurate. However, in general their use in predicting fricative behaviour is limited (Scully 1990), perhaps as a result of their over-simplification of the processes occurring within the tract. From such models nonetheless, Fant (1970) demonstrated that forward excitation would both excite the natural resonances in the tract, but also produce a number of anti-resonances, or zeros, although it could not be predicted where these would occur for a given tract configuration. These zeros would appear as troughs in the spectrum when not located near any poles, and a pole and zero in close proximity would tend to cancel each other out. Generally, many of the back-cavity poles are attenuated by zeros, giving spectra their characteristic 'broad' shape. It was also demonstrated that the forward excitation produces a characteristic zero at low frequencies. The characteristics of several fricatives were summarised by concluding that [f] has no, or a very high resonance frequency (due in part to the lack of a resonant cavity in front of the constriction). The [s] configuration displayed the properties of a high-pass filter with high cutoff frequency. The large resonant cavity in front of the [ʃ] constriction gave it a single resonant frequency lower than [s] or [f]. It seemed likely that multiple sources could exist, and Fant (1970) concluded his work by pointing out the importance of increased measurements of productions.

The experiments performed by Shadle (1990) revealed a significant amount of new information about forward production, with physical models incorporating flow obstructions, and with appropriate explanations for various findings. In addition to the zero found close to 0 Hz during forward excitation, an additional complex-conjugate pair of zeros are generated at a distance inversely proportional to the distance between the constriction, and the sound source. It was also noted that since this distance is usually very short, a small change (say, ~1mm) will move the first free zero by a significant amount (possibly several hundred hertz). In addition, the distance between source and constriction was found to be inversely proportional to the amplitude of the radiated sound.

Nevertheless, the considerable variability in the observed spectral characteristics of fricatives is still not accounted for by current models (Stevens 1998).

### 2.1.3 A partially-known stochastic process

A sustained voiceless fricative sound — for example [ʃʃʃ] — is characterised by random turbulent airflow that generates acoustical noise. The resulting far-field signal can be considered as a stationary stochastic process. This implies that, for a *fixed* tract configuration and lung pressure, time series (and hence the frequency spectrum) produced by the system will be different over any two time intervals, despite having identical underlying statistical properties. Such stochastic systems can only be effectively described qualitatively in terms of estimated statistics, rather than quantitatively.

This stochastic process can be modelled as a noise source acting on a pole-zero filter, and so the spectral 'peaks' and 'troughs' can be described by a number of poles and zeros. It seems likely that the positions of these poles and zeros are significant to the production of fricative sounds, albeit via a highly nonlinear relationship. The frequency location of poles, and hence peaks in the spectrum are generally determined by resonant frequencies in the tract. The positions of zeros, and hence troughs in the spectrum are *determined in large part by the precise position of the forward noise source*, although the interaction of source position and zero location is hard to deduce from the speech signal alone. Variations of air pressure will lead to changes in turbulent characteristics, and so intricate relationships between fricative intensity and the spectrum are likely to exist.

To clarify some facts that are known about turbulent-type production, it may be appropriate to briefly consider a first-order approximation of a multiple-forward-source tract system. It is known that a fixed tract will exhibit fixed poles intrinsic to its shape. Next, consider a quadrupole or dipole source positioned somewhere in the tract. This source will excite the poles in the tract function, and if it is in a forward position in the tract, will also generate a particular configuration of zeros. The overall output spectrum of the system so far will be approximately represented by the product of source spectrum, tract poles, and zeros caused by the forward position of the source. However, if the particle velocity is changed (as a result of an increase in overpressure) then it can be expected that the source spectrum will change, and so the source spectrum is velocity-dependent. It has also been observed that such a change in flow will also result in a positional change of the source, resulting in a change in the zeros generated by this source. It should also be considered that within the system, a number of both quadrupole and dipole sources will be present. A further complication therefore arises, since the flow velocity in one part of the tract will in general be dependent upon configurations posterior in the tract.

Multiple zeros at a range of frequencies are introduced for a given configuration with forward excitations. The problem of estimating the source distributions from the resulting acoustical signal becomes much more complex, and the solution is no longer unique. Further, the interactions of multiple sources is overwhelmingly complicated. It rapidly becomes clear why capturing specific articulatory information from the output spectrum is such a difficult task.

Despite expectations that such systems are highly complex, it is entirely possible that they may also exhibit statistical covariances. Good analysis tools should try and make use of this data. However, perhaps as a result of inadequate treatment of fricative signals, the reverse is often implemented, and the contributions of features of known importance are suppressed. In particular, care should be taken over the degree of frequency-smoothing that is performed (discussed in §2.2.2.3), which can degrade spectral peaks and troughs. The use of spectral moments as a broad descriptive metric is discussed in §2.3 and Chapter 5.

The voiceless fricatives appearing in connected speech, while exhibiting some of the stochastic characteristics of sustained voiceless fricatives, introduce multiple nonstationary elements (Scully and Allwood 1985; Scully 1990). For example, the tract cross-sectional area varies over time in fricatives in /VFV/ context (Scully, Grabe-Georges, and Castelli 1992). Also, it is likely that the forward noise source location moves during production, and this will be reflected by changes in amplitude, moving zeros and hence troughs (Shadle and Scully 1995); furthermore, these changes are affected in different ways by specific vowel context: they cite an example of a subject whose [s] productions were found to be most significantly affected in /usu/ context. The signal must therefore be considered a nonstationary stochastic process.

It has been suggested many times in the literature that the central region of a spoken fricative can be considered reasonably stationary. However, this seems to be a loose assumption and any nonstationary regions in real speech that are treated as stationary will produce increasingly inaccurate results as larger time-series data segments are relied upon. Unfortunately, this 'brief stationarity' introduces problems in stochastic frequency analysis, and is discussed in §2.2.2.2; the effects resulting from erroneously assuming the signal to be stationary will be demonstrated later in §4.2.

Compounding the complexity of the process is the evidence that, on repeating a spoken fricative, a single speaker will often vary the precise manner in which the fricative is generated (Scully and Allwood 1985). This means that our nonstationary stochastic process is also not ergodic. This has implications for ensemble-averaging techniques, and will be discussed in §2.2.2.1. Additionally, since the productions across speaker vary significantly (Scully 1990), but generally with invariable perceptual results, it is clear that methods of measuring acoustical differences across productions are needed.

These facts mean that in order to analyse fricative signals, great care must be taken. Indeed, a rather fragile set of additional constraints and assumptions must be made if any measurements are to be taken at all, and these are discussed in §2.2.

The fricative production system is based on one or more noise-sources, whose location may vary across productions, and which excite a system of chambers that produce anti-resonances as well as resonances (that may also vary significantly across productions), and whose precise locations will help describe the system. It has been found that knowledge of these precise source properties is highly significant if either the source, or the overall system is to be modelled. Additionally, even very 'simple' articulation adjustments result in complex acoustic pattern changes (Scully et al. 1992). Nevertheless, we can expect that correlations connecting these variables exist, and if measurable, would be invaluable in our understanding of fricative production (Scully, Castelli, Brearley, and Shirt 1992):

> The multiplicity of acoustic effects resulting from the articulatory actions are not independent of each other however: the acoustic sources are linked by the unified aerodynamic system of the whole respiratory tract; the actions which control the sources also determine changes in formant frequencies and bandwidths. Co-varying acoustic pattern changes should be expected, governed by the aerodynamic and acoustic processes of speech production working on the particular articulatory scheme chosen by the speaker. It seems likely that any or all of these acoustic patterns, including their pattern of co-occurrences, may be useful to listeners and may be important when characterising and modelling sequences containing fricative consonants.

Measuring the covariances between configurations across productions may lead to increased insight as to the nature of fricative production, and this is explored in Chapter 6.

One of the most important tasks therefore becomes that of making precise measurements of productions of the system, and rather than ignoring knowledge about the production procedures, making as much use as possible of them, and carefully observing the typical changes that occur between productions.

## 2.2 Nonparametric spectral estimation

This section serves as a reminder of some key principles concerning the spectral estimation of stochastic processes, as well as introducing a few modern methods that have recently been developed. A thorough treatment of Fourier and statistical theory is not given here, since these topics are well-covered elsewhere in the literature (e.g. see Bendat and Piersol 1986; Percival and Walden 1993). Familiarity with fundamental signal processing principles such as Nyquist criterion, linear filters, and so on, is assumed. No rigorous proofs are undertaken, but the procedures of the above authors are followed.

It must be noted that many well founded principles of spectral estimation of stochastic processes are often overlooked in the area of speech analysis, and hence §2.2.1 starts us off with a refresher of the most important aspects, including Fourier interpretation of stochastic processes, statistical errors in estimates, the periodogram, and data tapers. Section 2.2.2 reviews some established procedures in suitable treatment of fricative signals, ensemble averaging, time averaging, and frequency smoothing, which are surprisingly often absent from many fricative analysis research publications.

Section 2.2.3 introduces multitaper spectral analysis, a valuable new tool available to speech science, although not often used. Again, rigourous analysis of this methodology is present in the literature. In particular, the reader is referred to Percival and Walden (1993), and Thomson (2000). Many of the key principles in these works are presented in this section, since they are of significant importance in the spectral analysis of fricatives.

### 2.2.1 Principles

Before we begin describing some of the more advanced signal processing techniques required for proper treatment of fricative signals, it is necessary to consider the basic facts and quantities associated with spectral estimation of stochastic processes. Unless stated, a sampling frequency ($f_s = 1/\Delta t$) of 1 Hz can be assumed.

#### 2.2.1.1 Fourier methodology

Fourier theory outlines fundamental principles for describing time series data in terms of the amplitudes and frequencies of the cosine and sine waves that it is composed of. Generally, any given time series can be described in the following form:

$$x_t = \mu + \sum_{k=1}^{N/2} [A_k \cos(2\pi f_k t)) + B_k \sin(2\pi f_k t)] \tag{2.3}$$

where $\mu$ is the mean value of the discrete time series $x_t$, and $A_k$ and $B_k$ are the amplitudes of the cosine and sine components at frequencies $f_k$ of the time series $x_t$.

In the case of a stochastic process, the coefficients of the frequency components are interpreted as random variables, whose variances $\sigma_k^2 = E\left\{A_k^2\right\} = E\left\{B_k^2\right\}$ are of interest, and hence are the variables we are trying to estimate for the process. If we assume $\mu = 0$, then the ***discrete power spectral density function*** is defined

$$S(f_k) \equiv \sigma_k^2 \tag{2.4}$$

for $1 \leq k \leq N/2$. The power spectrum describes the contributions of energy around the frequencies $f_k$. The spectral density function for a process allows spectral properties of the process to be more easily interpreted than from time-series data. We can expect that tract resonances, and turbulence noise source shapes may show up as distinctive spectral features.

#### 2.2.1.2 Errors and statistical measures in random variable estimation

Since we can rarely hope to observe the complete *ensemble* of sample sequences of any given stochastic process, we can only hope to form an approximate estimate of the characteristics of the process. Such an estimate will inevitably be the subject of errors, and we now take a moment to compare the different types of error which we can expect to encounter.

The ***bias*** $b\{\cdot\}$ of an estimate $\hat{\zeta}$ of some variable $\zeta$ is the systematic error, defined

$$b\left\{\hat{\zeta}\right\} = E\left\{\hat{\zeta}\right\} - \zeta. \tag{2.5}$$

A good estimate will have reducing bias with increasing sample size.

The *random error* that exists when trying to measure a parameter, known as the *variance* var$\{\cdot\}$ of the estimate, is defined

$$\text{var}\left\{\hat{\zeta}\right\} = E\left\{(\hat{\zeta} - E\{\hat{\zeta}\})^2\right\} \tag{2.6}$$

and an ideal estimator will have decreasing estimate variance with increasing sample size.

An estimator whose bias and variance disappear as the number of samples under analysis approaches infinity, such that

$$\lim_{N \to \infty} E\left\{\hat{\zeta}\right\} = \zeta \tag{2.7}$$

is of course most desirable, and said to be *consistent.*

Finally, it is desirable that two estimators of uncorrelated parameters $\zeta_1$ and $\zeta_2$ of the process, are themselves uncorrelated, so that

$$\text{cov}\left\{\hat{\zeta}_1, \hat{\zeta}_2\right\} \approx 0 \tag{2.8}$$

where $\operatorname{cov}\{\cdot,\cdot\}$ is the covariance between two variables, defined as

$$\operatorname{cov}\{\zeta_1,\zeta_2\} \equiv E\{(\zeta_1 - E\{\zeta_1\})(\zeta_2 - E\{\zeta_2\})\}. \tag{2.9}$$

#### 2.2.1.3 The periodogram

The discrete Fourier transform of the discrete signal $x_{t,N}$, which is equal to zero when $t$ is outside $[1,N]$, is given by

$$\begin{aligned} G_N^{(p)}(f) &= \sum_{t=-\infty}^{\infty} x_{t,N}e^{-i2\pi ft} && (2.10)\\ &= \sum_{t=1}^{N} x_t e^{-i2\pi ft} && (2.11)\end{aligned}$$

where the parenthesised superscript is used to indicate the type of estimate under consideration (in this case, a 'p' for 'periodogram'). From (2.3), (2.4) and (2.11), it can be shown that the power spectral density function defined in terms of the Fourier transform, can be written as

$$S(f) \equiv \lim_{N\to\infty} E\left\{\frac{\left|G_N^{(p)}(f)\right|^2}{N}\right\}, \tag{2.12}$$

where $S(f)df$ is the expected contribution, over all possible realisations of the process, to the power from components with frequencies in the interval around $f$.

An appropriate approximation for a finite sample sequence would appear to be

$$\hat{S}^{(p)}(f) = \frac{\Delta t}{N}\left|\sum_{t=1}^{N} x_t e^{-i2\pi ft\Delta t}\right|^2, \tag{2.13}$$

where $\Delta t$ is the sampling period. This estimate is known as the *periodogram* of the power spectrum. Let us take a moment to consider the properties of this spectral estimate.

If $x_t$ is set to unity everywhere, the expected response for a DC signal can be observed. The periodogram estimate is equal to Fejér's kernel $\mathcal{F}(f)$, which is shown on the right of Figure 2.1. This well known response demonstrates that, for a single underlying frequency component, the response will be distributed between a number of 'lobes'. The main lobe, centred on the frequency under examination, contains the most energy, but much of the energy for this response is dispersed over the rest of the frequency range, in what are known as the 'sidelobes'. In fact, whatever sequence $x_t$ is set to, the expected power spectrum will be equal to the convolution of the sequence's actual power spectrum, with Fejér's kernel:

$$E\left\{\hat{S}^{(p)}(f)\right\} = \int_{-f(N)}^{f(N)} \mathcal{F}(f-f')S(f')df' \tag{2.14}$$

The redistribution of energy away from the frequencies at which that energy originates is known as 'leakage', and is undesirable, since it introduces bias, and significant spectral correlation, especially near prominent spectral peaks. That is, for any process that is not spectrally 'flat',

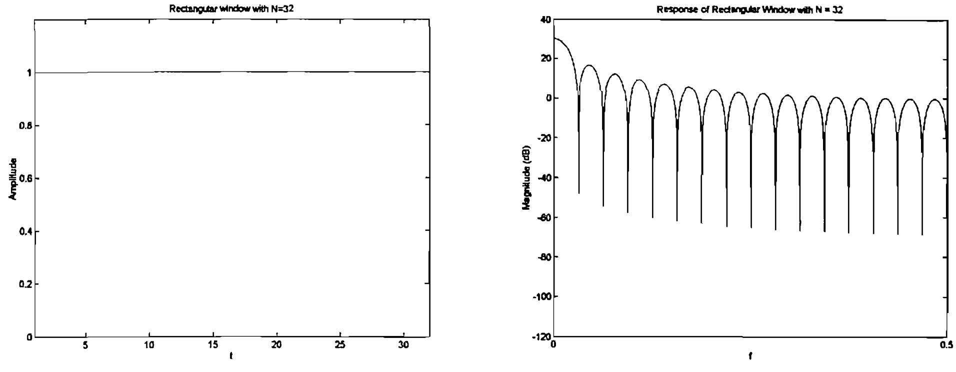


FIGURE 2.1: Rectangular time window ($N = 32$) (left) and frequency response, known as Fejér's kernel (right).

and at frequencies where the underlying power spectral density is low, the estimate will indicate much higher energy than actually exists.

The 'flatness' of a power spectrum can be described in terms of its *dynamic range*, simply the quantity

$$10\log_{10}\left(\frac{\max\{S(f)\}}{\min\{S(f)\}}\right) \tag{2.15}$$

Percival and Walden (1993) show that the properties of Fejér's kernel are such that, as N approaches infinity, the bias of the estimate vanishes. However, in practical terms, even with large time series, the bias can be significant. For example, for a spectrum with known dynamic range of around 60dB, when $N = 1024$ the nonlocal bias (i.e. far from the main spectral peaks) was shown to be of the order of 18dB.

In order to calculate the expected variance of the estimator, we first assume that the random variables $A_k$ and $B_k$ of the magnitudes of the cosine and sine components of the process have Gaussian distributions. Then, since the periodogram spectrum can be written as

$$\hat{S}_G^{(p)}(f_k) = A^2(f_k) + B^2(f_k), \tag{2.16}$$

and the sum of $\nu$ squared Gaussian variables has a chi-squared distribution with $\nu$ degrees of freedom,

$$\chi_\nu^2 = Y_1^2 + Y_2^2 + \ldots + Y_\nu^2, \tag{2.17}$$

then it is straightforward to show that, at any given frequency, the periodogram estimate is a chi-squared distribution with two degrees of freedom:

$$\hat{S}_G^{(p)}(f_k) = \frac{\sigma_k^2 \Delta t}{2}\chi_2^2 \tag{2.18}$$

for $0 \leq f_k \leq f_N$. Now, since $E\left\{\chi_\nu^2\right\} = \nu$ and $\text{var}\left\{\chi_\nu^2\right\} = 2\nu$, then it is trivial to show that

$$E\left\{\hat{S}_G^{(p)}(f_k)\right\} = \sigma_k^2 \Delta t = S_G(f_k) \tag{2.19}$$

that is, the mean value for the estimator is equal to the value being estimated, but on observation of the variance, the rather more disastrous result

$$\mathrm{var}\left\{\hat{S}_G^{(p)}(f_k)\right\} = \sigma_k^4(\Delta t)^2 = S_G^2(f_k) \tag{2.20}$$

for $0 \leq f_k \leq f_N$ is found. This result show that, irrespective of sample size $N$, the variance of the periodogram estimate is equal to the value that we are trying to estimate.

In most cases this error should be viewed as too large for this estimator to be of any use. However, spectral estimates with variance errors of this size are highly common in the fricative speech literature.

In summary, for any stochastic process whose underlying power spectrum is not flat, the periodogram estimate is the subject of large bias error near spectral peaks. Moreover, in regions of high energy, it has large variance.

#### 2.2.1.4 Data tapers

In order to reduce the leakage from the main lobe into the sidelobes, and hence the bias, the periodogram is invariably calculated under a data taper or window. Data tapers work by smoothing the extremities of the time series data to be analysed, which results in a reduction of the sidelobes of the response. A data taper that accomplishes this will of course lead to power spectral estimate with much lower bias, which is desirable.

For a given length of time series data, a discrete data taper $h_t$ takes the form of a number of weights that are used to pre-emphasise the time series data, prior to calculating the Fourier transform, and hence estimating the power spectrum.

$$\hat{S}^{(mp)}(f) = \Delta t \left| \sum_{t=1}^{N} h_t x_t e^{-i2\pi f t \Delta t} \right|^2 \tag{2.21}$$

where $\sum h_t \equiv 1$. This estimate is known as the modified periodogram, and commonly in speech analysis, Hanning or Hamming windows are used. A Hanning window of length $N = 32$ is shown on the left of Figure 2.2. It is necessary to briefly remind ourselves of the properties of these tapers.

The expected estimate calculated under a Hanning data taper is

$$E\left\{\hat{S}^{(dt)}(f)\right\} = \int_{-f(N)}^{f(N)} \mathcal{H}(f - f')S(f')df' \tag{2.22}$$

where $\mathcal{H}(f)$ is the power spectral response of the Hanning window, shown on the right of Figure 2.2. Note the reduction in magnitude of the sidelobes, but also the increase in width of the main lobe. This results in a decrease in the bias, and hence the estimate is invariably more appropriate. For further discussion on data tapers, their spectral responses, and estimate bias, see Priestley (1999) pp.556-574.

This estimate is commonly used in fricative analysis research. However, while it satisfactorily reduces bias, it does not tackle the large estimate variance problem discussed in §2.2.1.3.

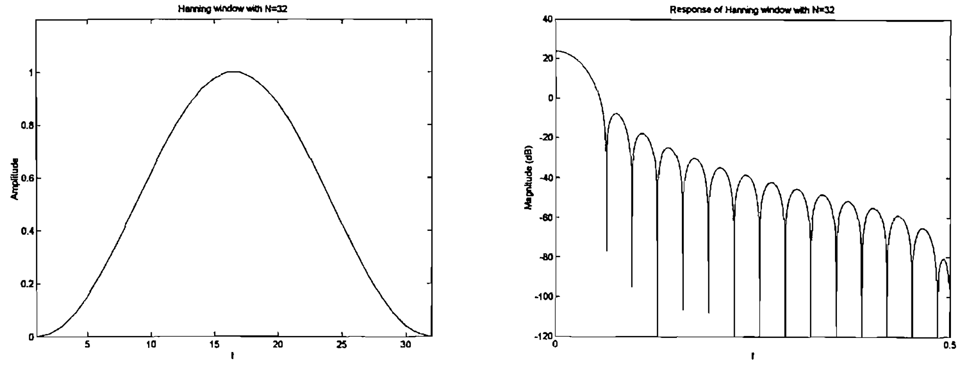


FIGURE 2.2: Hanning time window ($N = 32$) (left) and frequency response (right).

In fact, by introducing the data taper, much of the data towards the edges of the data window is almost 'discarded', as a result of the small weightings at these points. This has the effect of increasing the estimate variance further, by a factor of approximately two (Bendat and Piersol 1986).

That is, much of the fricative analysis literature makes use of estimates whose variance error is nearly twice as large as the underlying distribution itself. This may be one of the root causes of the difficulties and general lack of success to date, of characterisation methods that attempt to track spectral features (and hence, tract resonances) during fricative production. The general popularity of spectral moments in the face of such spectral estimators becomes clearer, as discussed in §2.3.

To properly deal with the large estimate variance, some form of averaging must take place.

## 2.2.2 Averaging methods

In the last section it was shown how data tapers can be used to reduce the bias error of the spectral estimator. For our estimator to be of any practical use however, it is necessary for the large variance of the estimate to be dealt with. In order to reduce it, some assumption about the underlying process must be made, and hold true, so that some form of averaging process can be used to generate a more accurate estimate.

### 2.2.2.1 Ensemble-averaging

As previously discussed, the analysis of random data requires some form of averaging if consistent estimates of parameters describing that data are to be obtained.

If the process can be considered ergodic (i.e. if the statistical properties of the process are independent of sample sequence), the variance of the estimate of the power spectrum can be reduced using an *ensemble* of sample sequences. As long as the assumption can be made that the statistical properties of the process are independent of sample sequence, consistent estimates of these characteristics can be calculated by averaging over a number of sample sequences.

A consistent estimate of the power spectrum of a stationary random process can be calculated over $N_d$ sample sequences by

$$\hat{S}^{(e)}(f) = \frac{1}{N_d} \sum_{u=1}^{N_d} \hat{S}_u^{(mp)}(f) \tag{2.23}$$

where $\hat{S}_u^{(mp)}(f)$ is the spectral estimate of sample sequence $u$. Unfortunately, the feasibility of using this method in practice means we are often limited to a small number of sample sequences over which to form an average estimate. Nevertheless, it is straightforward to show that the variance of the estimate is reduced to

$$\operatorname{var}\left\{\hat{S}^{(e)}(f)\right\} = \frac{S(f)}{N_d}, \tag{2.24}$$

while the expected form of the estimator remains as (2.22). From this result it is clear that, for any $N_d$ independent sample sequences of the stochastic process, the variance of the new estimate is reduced by a factor of $N_d$.

However, the success of ensemble averaging relies upon the assumption that the underlying process is ergodic. If it is desired that the power spectra of individual sample sequences from some non-ergodic stochastic process be analysed, this method is of limited use.

These facts give rise to problems in attempting to form consistent power spectral density estimates using ensemble averaging techniques. A method of consistent spectral estimation is therefore desired that does not require an ensemble.

#### 2.2.2.2 Time-averaging

A discrete stochastic process $\{x_t\}$ is said to be weakly stationary if $E\{x_t\} = \mu$, and $\operatorname{cov}\{x_t, x_{t+\tau}\} = s_\tau$ for $\tau = 0, \pm 1, \pm 2, \ldots$, where $\mu, s_\tau$ are finite constants independent of $t$. Another measure of stationarity is whether the statistical moments of the process are independent of time. If a process is stationary, the variance of the estimate of the spectrum can be reduced by averaging the power spectra from several independent windows of time series data.

The method formalised by Welch (1967) splits the sample sequence of length $N$ into $N_B$ smaller subsequences, each of length $N_S$. Each of the subsequences consist of samples $l, l+1, l+2, \ldots, l+N_S-1$. The power spectrum of each subsequence (which of course has reduced spectral resolution) is then estimated by

$$\hat{S}_l^{(mp)}(f) \equiv \Delta t \left| \sum_{t=1}^{N_S} h_t x_{t+l-1} e^{-i2\pi f t \Delta t} \right|^2 \tag{2.25}$$

The estimated power spectra of several subsequences can be combined in order to reduce the variance of the estimate. However, it is not necessary to average over all subsequences, since very little information is gained by moving between neighbouring subsequences. Rather, the data window advances through the sample process by some constant $N_a$, where $0 < N_a < N_S$

usually. That is, Welch's segmented averaging method of spectral estimation is

$$\hat{S}^{(W)}(f) \equiv \frac{1}{N_B} \sum_{q=0}^{N_B-1} \hat{S}^{(d)}_{qN_a+1}(f). \tag{2.26}$$

In the case where the subsequences used in (2.26) are completely non-overlapping, the variance of the spectrum will again be of the form (2.24), where $N_d \approx N_B/2$. However, it is more common in the engineering literature for the windows to overlap by around 50% (corresponding to a value of $N_a \approx N_B/2$). This allows some of the data that would have been attenuated by the data taper to be recovered. In this case of 50% taper overlap, it has been shown that the effective number of degrees of freedom is approximately

$$\nu \approx \frac{36N_B^2}{19N_B - 1}. \tag{2.27}$$

which is equivalent to the factor by which the variance of the estimate will be reduced.

This result indicates much better accuracy for power spectral estimation using time averaging. The expected form of the response remains as (2.22), suggesting no deterioration of the spectral bias, although the resolution is reduced by a factor of $N_B$. However, we must bear in mind that this method relies heavily on the signal being stationary. Treating a random signal with nonstationary properties as stationary will clearly distort the spectral density estimate, since the frequency representation of a nonstationary signal is ill-defined.

#### 2.2.2.3 Frequency smoothing

If the process under examination is neither stationary, nor ergodic, then averaging the spectral density estimate over a small interval of frequencies is an alternative method of reducing the variance of the estimate, so long as the underlying spectrum is smooth. When averaging over the frequency interval, the spectral density estimates at each of the nearby frequencies can be weighted, using a *spectral* window, $W(f)$. The frequency smoothed estimate is hence given by:

$$\hat{S}^{(fs)}(f) \equiv \int_{-f(N)}^{f(N)} W(f-\phi)\hat{S}^{(d)}(\phi)d\phi \tag{2.28}$$

Naturally, the choice of spectral window can be optimised if detailed knowledge of the underlying process is known. Typically in speech analysis, the *Daniell* spectral window may be used (simply an even weighting over the $M$ neighbouring frequencies):

$$\mathcal{D}_M(f_k) = \begin{cases} \frac{1}{M+1} & \frac{k-M}{2} \leq k \leq \frac{k+M}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{2.29}$$

and we will only consider the properties of this spectral window here. A multitude of different spectral windows are in existence; for a more detailed discussion of some of these, see Percival and Walden (1993), and Priestley (1999).

Figure 2.3 shows four example responses, after a Daniell window (of increasing values of $M$) has been used to smooth the responses calculated using a Hanning window, shown in Figure 2.2. The

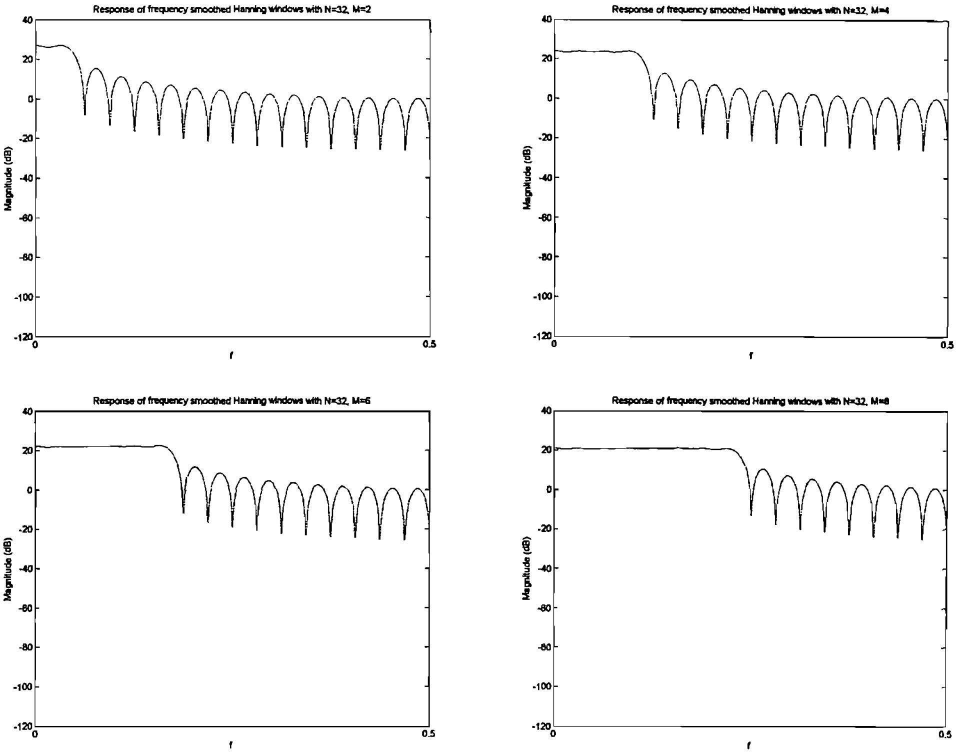


FIGURE 2.3: Frequency response of frequency-smoothed spectra for $M = 2$ (top-left), $M = 4$ (top-right), $M = 6$ (bottom-left) and $M = 8$ (bottom-right), using Hanning window ($N = 32$).

first thing to be noticed is the flat, wider response of the main lobe. This highlights the frequency range over which the averaging takes place, and also over which the assumptions of flatness must hold. The frequency smoothed estimate will decrease the variance of estimate, so long as the underlying spectrum is smooth. This process will therefore have the effect of greatly increasing the spectral correlation between nearby frequency values, and also significantly increase the local bias, an expression for which is (as given by Percival and Walden (1993)):

$$b_w(f) \approx \frac{S''(f)}{24}\beta_w^2 \tag{2.30}$$

where

$$\beta_w \equiv \left(12\int_{-f_{(N)}}^{f_{(N)}} f^2 W(f)df\right)^{1/2} \tag{2.31}$$

showing that the local bias is influenced by both non-smooth features near the frequency region of interest, as well as the size of the spectral window. Nevertheless, the overall reduction in the variance of the resulting spectral estimate is again of the form (2.24), where $N_d = M/2$.

Great attention should also be paid to the considerable increase in the bias of this estimate with increasing $M$, since the sidelobes of this response have grown significantly with the frequency smoothing operator. In fact, this new bias, resulting from the spectral smoothing operator, is most effectively reduced by choosing a more appropriate data taper under which to weight the time series data, before any frequency smoothing operations are performed. One such set of data tapers are discussed now in §2.2.3.

### 2.2.3 Introduction to multitaper analysis

Data tapers have generally been designed with some specific property in mind, in order to suit some specific task. For example, to view the formants in vowel speech spectra, the Hanning and Hamming windows commonly have the most suitable properties. When used for the power spectral estimation of stochastic processes however, their use should be reevaluated.

The development of the multitaper methodology for good spectral estimation of stochastic processes has been significantly contributed to by Slepian (1978) and Thomson (2000). We now take a moment to describe the rationale behind the multitaper methodology, following the work by Percival and Walden (1993).

Consider some sample sequence. In order to estimate the spectral density of this sequence, without significant spectral leakage, the data must be weighted in a suitable manner. However, by weighting with a smooth window that reduces sidelobe leakage, some of the time-series data are invariably lost, leading to an increase in the variance of the estimate. If these data could be recovered in some way, the variance of the estimate could be reduced.

Consider then, a set of orthogonal data tapers, which could each be used to estimate the spectrum of a different (orthogonal) portion of the sample sequence data. If each of the data tapers has a good spectral response (i.e. one with small sidelobe leakage), then the spectral estimates using each of the tapers will have small bias, although the variance of each estimate will be large.

However, due to the linearity of the Fourier transform, the responses of our orthogonal set of time windows will themselves be orthogonal. In this case, the orthogonal spectral estimates can be averaged to produce a new estimate with reduced variance.

The *concentration* of a time signal, or data taper, can be defined:

$$\alpha^2(T) \equiv \frac{\int_{-T/2}^{T/2} |h(t)|^2 \, dt}{\int_{-\infty}^{\infty} |h(t)|^2 \, dt} \tag{2.32}$$

which is the fraction of the taper's total energy, in the time interval $T$ centred around 0. A similar expression exists for the frequency concentration of a taper's spectral response:

$$\beta^2(W) \equiv \frac{\int_{-W}^{W} |H(f)|^2 \, df}{\int_{-\infty}^{\infty} |H(f)|^2 \, df} \tag{2.33}$$

The orthogonal set of tapers should attempt to maximise (2.32), while restricting (2.33) to some predefined 'acceptable' bandwidth $2W$. It has been shown that the solutions to this maximisation problem take the form of an orthogonal set of eigenfunctions $\psi_\kappa(\cdot; c)$ (where $c \equiv \pi WT$), corresponding to the data tapers themselves, known as *prolate spheroidal sequences*, or *Slepian* sequences in recognition of his significant contributions. Each of these eigenfunctions has a corresponding eigenvalue $\lambda_\kappa$, proportional to the energy of each taper.

Four Slepian sequences computed for $M = 4$ and $N = 32$ are shown on the left of Figure 2.4, with their corresponding responses on the right. As can be seen, the first Slepian sequence is rather similar to the familiar Hanning window: the data at the ends of the sequence are attenuated, and the sequence is smoothly introduced in order to minimise sidelobe leakage. As can be seen from its corresponding frequency spectrum on the right, the sidelobes indeed drop rapidly from the main peak. The main peak is $2Wf_s$ wide, as defined by (2.33).

The second Slepian sequence is rather different to most tapers that are usually found in the speech literature. Peculiarly, the window goes negative for some duration. This would not be used to calculate deterministic frequency spectra, since it would introduce significant phase distortion. However, the phase of a stochastic process is of no importance. Again, the response consists of a large primary lobe, with rapidly attenuated sidelobes. Of interest is that the width of the main lobe is the same as all the other responses (as determined by (2.33)), although the peak of the main lobe is shifted slightly. This corresponds to the expected orthogonal responses, which, when combined, will work to reduce the estimate.

Each Slepian taper can be used to form an estimate:

$$\hat{S}_{\kappa}^{(mt)}(f) \equiv \Delta t \left| \sum_{t=1}^{N} \psi_{t,\kappa} x_t e^{-i2\pi f t \Delta t} \right|^2 . \tag{2.34}$$

Each of these estimates will have good sidelobe characteristics (because (2.32) has been minimised), but a large variance (more so since the tapering operation reduces the amount of data present with which each estimate is formed).

However, since these spectral estimates are orthogonal to each other, their average response

$$\bar{S}^{(mt)}(f) \equiv \frac{\sum_{\kappa=0}^{M-1} \lambda_{\kappa} \hat{S}_{\kappa}^{(mt)}(f)}{\sum_{\kappa=0}^{M-1} \lambda_{\kappa}} \tag{2.35}$$

will have a variance reduced by a factor of $M$, as more of the time series data is incorporated into the estimate. Additionally, since the bandwidth has been restricted by (2.33), we can expect that the *combined* response of $\bar{S}^{(mt)}(f)$ will have optimal bias properties.

For a given application, the acceptable resolution bandwidth factor $W$ (where $0 < W \leq 1/4$) will determine the family of Slepian sequences to be used in an analysis. The time-frequency resolution tradeoff is still controlled by the length $N$ of windows that we choose to use, and the quantity $NW$ is known as the *bandwidth product*. As $W$ is increased, the width of the main lobe increases, but so does the number of Slepian windows $M \approx 2NW$ that can be used in the analysis, whilst still limiting the resolution. Typical values for the bandwidth product are $4 \leq NW \leq 6$ (Thomson 2000).

Examples of combined responses for this estimate are shown in Figure 2.5, for increasing $M$. The most significant difference between these responses, and those of frequency smoothed Hanning estimates in Figure 2.3, is that, due to the orthogonality of the Slepian sequences, the bias near to the main peak does not accumulate, whereas for the Hanning windows, it does. The importance of this is that the multitaper estimate can be expected to have significantly reduced bias near steep spectral peaks (and troughs), especially for small $N$, making it an excellent choice for spectral estimation over short intervals of slowly non-stationary stochastic processes.

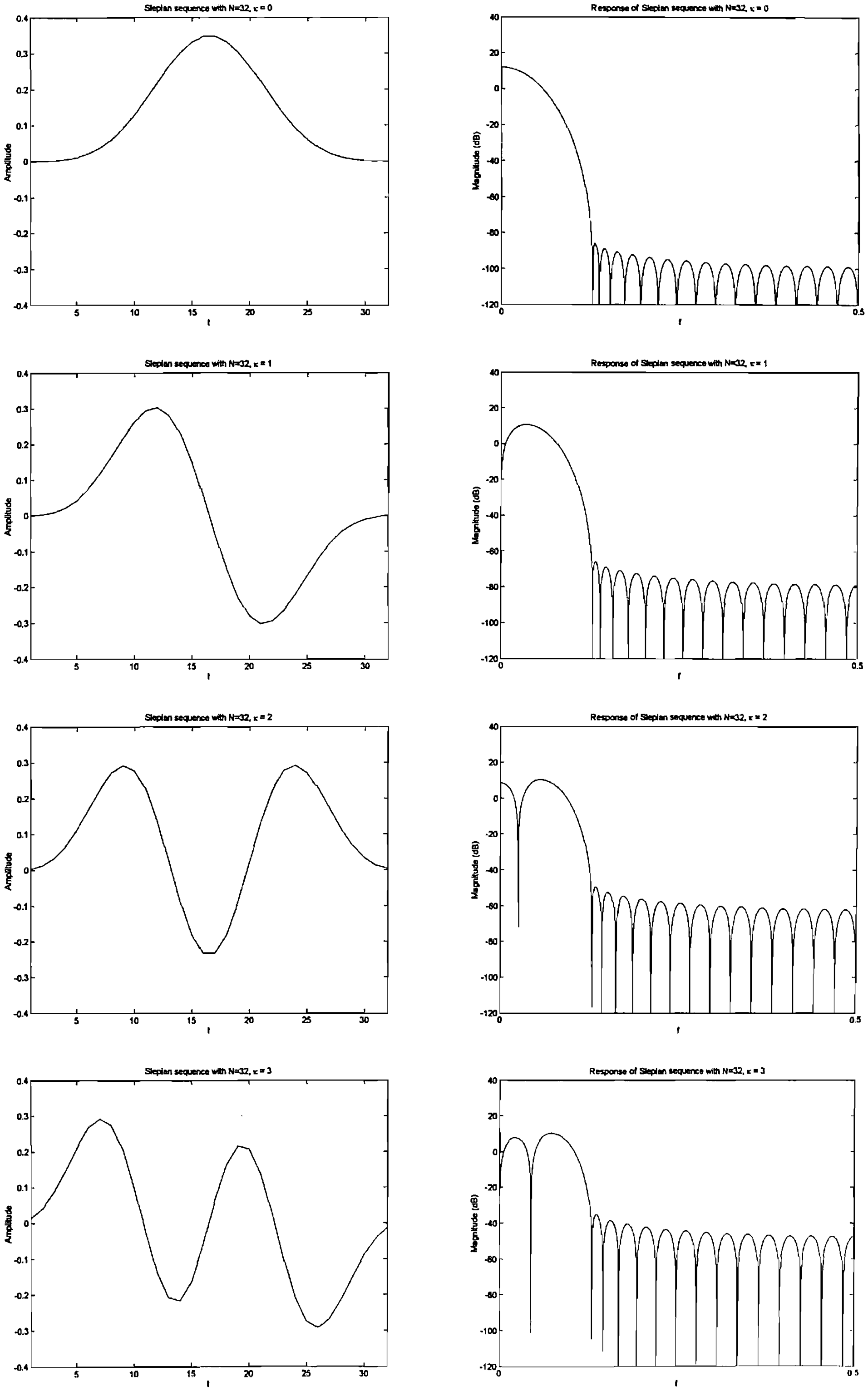


FIGURE 2.4: First (top) to fourth (bottom) Slepian tapers ($N = 32$) (left), and frequency responses (right).

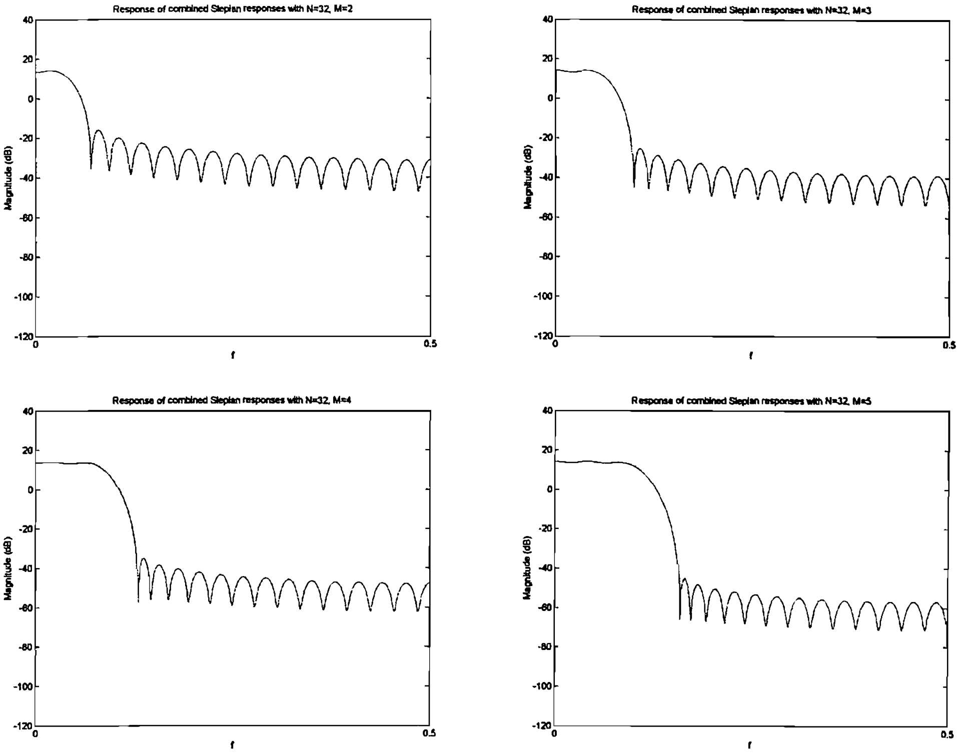


FIGURE 2.5: Frequency responses of multitaper spectra (N=32) for $M = 2$ and $M = 3$ (top), $M = 4$ and $M = 5$ (bottom).

### 2.2.4 Summary

It has been shown that, in order to reduce the bias of a spectral estimator, the data must be suitably tapered. However, in doing so, the variance of the estimate increases to almost twice the magnitude of the underlying spectrum itself.

In order to reduce this variance, some form of averaging must take place. In some fields, ensemble averaging is a feasible method of obtaining independent sample sequences from which to estimate the process spectrum. However, in the field of speech production, the practical problems related to this method are significant. It is well known that the production by a single speaker of identical words in identical contexts and situations will be produced with varying manner on each attempt. The mechanisms governing these changes in manner is not understood, and so must be treated as having a random element.

Time averaging is also a suitable method for reducing the variance of the estimate, provided the underlying process is stationary. However, in speech production, it is known that the articulators within the tract are constantly on the move in order to form the next phoneme. It seems unlikely that fricative production can be considered stationary enough for time-averaging techniques to work without distorting the true picture.

If fricative production is assumed to be neither ergodic, nor stationary, then only short sample sequences of time-series data may be used to estimate the spectrum. For short sample sequences, frequency smoothing operations will reduce the variance of the estimate, but also introduce

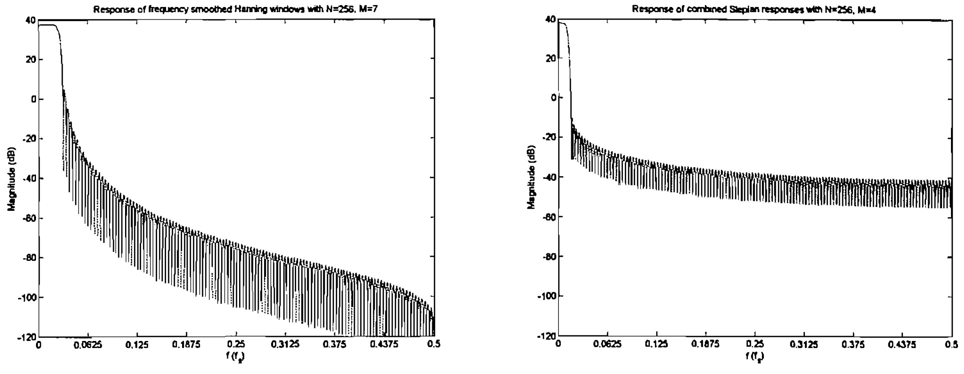


FIGURE 2.6: Comparison of frequency smoothed ($M = 7$) (left), and multitaper ($M = 4$) (right) responses for $N = 256$.

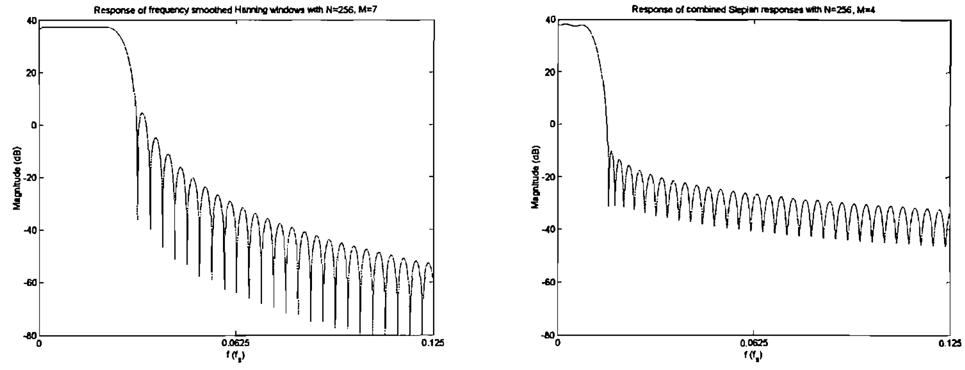


FIGURE 2.7: Comparison of frequency smoothed ($M = 7$) (left), and multitaper ($M = 4$) (right) responses for $N = 256$, magnified view.

significant bias to the spectral estimate.

Multitaper analysis provides an optimal way of reducing the bias of spectral estimates calculated over short intervals of sample sequence data. It is therefore highly likely that it is one of the most accurate methods available for the spectral examination of fricative production. However, the method has rarely been used for fricative analysis (see Blacklock and Shadle (2003)).

Figure 2.6 serves as a comparison of the response of Daniell frequency smoothing using Hanning data tapers (left), and of multitaper estimation. These two responses correspond to estimates that should have similar variance. Figure 2.7 highlights the region in which the multitaper's superior bias reduction can be viewed.

Demonstrations using speech signals of the differences between the different methods of reducing spectral variance and bias errors are given in Chapter 4.

## 2.3 Properties of spectral moments

Moments are used as a way of describing a given energy distribution with a small number of parameters. Specifically, the distribution is modelled as some deviation from a normal curve.

The first moment describes the mean energy location, or *centroid* of the distribution. The second moment describes the *spread* of the energy; that is, a distribution with most energy occurring near its centroid will have a smaller second moment than a distribution that has energy more evenly distributed across the range. The first two moments completely describe a normal distribution.

For non-normal distributions, higher order moments contain more information pertaining to the shape. The third moment, the *skewness*, is a measure of unevenness in energy distribution around the mean: a positive result indicates most energy is to the right of the mean, a negative result indicating most energy lies to the left of the mean. The fourth moment, known as *kurtosis* or *peakedness* is a measure of energy concentration in the immediate vicinity of the centroid. Higher order moments can be calculated for distributions, but these become increasingly abstract in terms of what they represent visually. The higher the order of moments used to describe the distribution, the greater the distribution can deviate from a normal curve. An infinite number of moments are required to describe any arbitrary shape.

Spectral moments then, apply these principles to spectra that can loosely be described as approximately-normal in shape. In order to do this, an arbitrary set of frequency and amplitude scales must be selected over which the analysis is to be performed. Since the moments of a distribution will be highly sensitive to whichever set of scales are used, it is important to take care in consideration.

### 2.3.1 Principles

Consider a discrete variate $X$ that takes the values $X_1, X_2 \ldots X_N$. Each of these values occurs with respective likelihoods $p(X_1), p(X_2) \ldots p(X_N)$, where $\sum p(X_n) = 1$. Then the *raw moments* of $X$ are defined (Kenney and Keeping 1964)

$$m'_r = \sum_{n=1}^{N} X_n^r \, p(X_n), \tag{2.36}$$

where $r$ is the moment order. The value $m'_1$ is known as the *mean* of the distribution.

*Central moments* are taken about the mean $\bar{X} = m'_1$

$$m_r = \sum_{n=1}^{N} (X_n - \bar{X})^r p(X_n). \tag{2.37}$$

The value $m_2$ is commonly known as the *variance* of the distribution. Furthermore, *standardised moments* may be calculated by normalising with respect to the function standard deviation $\sigma_X = \sqrt{m_2}$

$$\alpha_r = \sum_{n=1}^{N} \left( \frac{X_n - \bar{X}}{\sigma_X} \right)^r p(X_n). \tag{2.38}$$

The value $\gamma_1 = \alpha_3$ is commonly known as the *skewness*, while $\alpha_4$ is called the *kurtosis* of the distribution. Since the value of $\alpha_4$ for a normal distribution is equal to 3, the *excess of kurtosis* $\gamma_2 = \alpha_4 - 3$ is more commonly used.

Commonly, only the first four moments are considered, since these represent changes in the shape of the distribution that are clearly visible. Although a finite (and small) set of moments will always describe an infinite number of distributions, we can optimise the chances of 'recovering' the original distribution. In order to do this, it must be ensured that the distribution to be parameterised is well-suited to a 4-moment distribution, prior to calculation of moments. That is, for the spectral moment parameterisation method to work effectively, the model must hold well. Let us now remind ourselves of the meaning of the third and fourth order moments.

Briefly, the skewness (third order moment) of a distribution describes the asymmetry around the mean, particularly of the tail ends of a distribution. A positive value for the standard skewness of a distribution indicates that the tail is larger at values above the mean, and conversely, negative values represent greater probabilities of events below the mean.

Kurtosis (the fourth order moment) is a measure of the "peakedness" of a smooth distribution. A flat distribution will have a standard kurtosis value close to zero, while one which has greater probability of values near its mean will have a large standard kurtosis.

Of course we are free to consider higher order moments for the purposes of characterisation. Statistical methods can be used as a basis for determining the significance of the various moment measures.

### 2.3.2 Definitions

One of the earliest uses of the first spectral moment, usually referred to as the 'centre of gravity', was by Strevens (1960) to describe the distribution of energy in fricative power spectra. Although no specific values were calculated, Strevens reported clear visual differences in energy distribution for the different fricatives, and used the analogy of a spectral 'centre of gravity' to describe these patterns.

More specific measures of the 'centre of gravity' of fricatives were performed by Weinstein et al. (1975). Spectra were considered over a 0–5-kHz frequency range. If $S(f)$ represents the spectral amplitude, the centre of gravity for a given frame is given by $f_c = k_c\Delta f$, where $k_c$ is the largest integer for which

$$\frac{\sum_{k=k_c}^{k=127} S(k\Delta f)}{\sum_{k=0}^{k=k_c-1} S(k\Delta f)} \geq \theta_c.$$

where $\Delta f = 5000/128$. The quantity $\theta_c$ is specified in the text as having a value of $\frac{1}{2}$, but presumably this should be unity. The values of $f_c$ for five consecutive frames in the centre of the fricative are averaged together to form a single estimate of the 'centre of gravity' for that fricative. No further information is supplied about scales.

Jassem (1979) studied the spectra of fricatives over a total frequency range from 0–8 kHz. When compiling various quantitative features for use in multivariate analysis, the frequency range was divided equally into two and three 'fragments', and only the 'centre of gravity' of each of these fragments was calculated. Few other details about the calculation are given.

Forrest, Weismer, Milenkovic, and Dougall (1988) continued to develop the idea of calculating the centre of gravity by including higher order moments. The data were bandpass filtered to

70 Hz–10 kHz, and sampled at 20 kHz. The fricative portions of the speech signal were selected by hand. 20-ms data-windows (corresponding to 400 sample-points) were used in the analysis. A 400-point Hamming window was applied on each data-window, before zero-padding to create a new 512-point data-window. A 512-point FFT of this data window was calculated ($X(k)$), and the normalised power spectrum designated

$$p(k) = \frac{|X(k)|^2}{\sum_{n=0}^{256} |X(n)|^2} \tag{2.39}$$

for $k$ in $[1, 256]$. The first four moments were then defined:

$$\begin{aligned} L_1 &= \sum_{k=1}^{256} f_k p(k) \\ L_n &= \sum_{k=1}^{256} (f_k - L_1)^n p(k) \text{ for } n = \{2,3,4\}, \end{aligned}$$

where $f_k = k\Delta f$ and $\Delta f = 10000/256$. Further, the coefficients of skewness and kurtosis are defined $l_3 = L_3/\sqrt{L_2^3}$ and $l_4 = \left(L_4/L_2^2\right) - 3$ respectively. Additionally, a second method of moment acquisition is defined using a *Bark* frequency scale, as defined by Syrdal and Gopal (1986), whereby the power spectrum on linear frequency scale is mapped to the Bark scale, and weighted accordingly before moments are calculated. In all calculations then, it appears a linear power scale was used (i.e. corresponding to values of $|X_k|^2$, not $20\log_{10}|X_k|$).

In a further investigation into the capabilities of spectral moments, Jongman, Wayland, and Wong (2000) considered spectra up to 11 kHz of speech sampled at 22kHz (after lowpass filtering to 11 kHz). Data-windows of 40-ms duration (representing 880 data-points) from the centre of fricatives were first weighted with 40-ms Hamming windows, followed by fast Fourier transform (FFT) calculation. It is not stated whether the data were padded or clipped, but since it is declared that a higher frequency resolution than the Forrest et al. study was obtained, it is reasonable to assume that the data were zero-padded to 1024 points before a 1024-point FFT was calculated. The procedure given by Forrest, Weismer, Milenkovic, and Dougall (1988) was then followed: linear and Bark-scale frequency ranges were analysed, but only linear power scales were apparently used.

### 2.3.3 Spectrum frequency range selection

Historically, speech spectra have been considered up to typically 8–10 kHz, sometimes because it has been suggested that information at higher frequencies is superfluous to the task of speech discrimination, but in other cases simply as a result of limitations in recording equipment (e.g. Strevens 1960). However, since the ultimate aim is to describe the *production* mechanisms by the acoustical signals, it seems appropriate to consider a range that includes as much of the *produced* spectrum as possible.

Another factor that should be considered when determining an appropriate frequency range over which spectral moments can be calculated, is that it should result in a distribution that adequately fits the model: namely that it should be approximately-normal, and the number of moments being used should adequately describe the degree of deviation from a normal curve, and hence capture well its overall shape. If the number of moments being used is insufficient for describing the distribution, then the metric becomes inaccurate. If these errors are large, then significantly different distributions may produce similar sets of moments, which of course greatly limits their use.

Since normal distributions extend to infinity, but only have an asymptotically small amount of energy in these tails, the frequency range we select over which to analyse our spectrum should capture a spectral shape that approaches "zero" near the edges. In addition, only a single main "peak" should be contained within the frequency range, since multiple peaks will not be well described by the first four moments alone.

### 2.3.4 Spectrum magnitude scale selection

The first four moments describe a shape that is rather distinctively bell-shaped, if a little lopsided. It may therefore be necessary to adjust the magnitude scale of our power spectrum in order to obtain this characteristic, so that the moments act in a more complete manner: the distribution of maximum likelihood described by any set of the first four moments will always be approximately Gaussian.

Because speech sounds consist of a very wide range of energy intensities, power spectra often span several orders of magnitude at different frequencies in a single sound. The result is that power spectra on linear magnitude scale have a very "spiky" appearance, often rendering them difficult to read. A common practice is to plot power spectra on a logarithmic scale, often in decibels (dB). This has the effect of reducing the amplitude range to within readable limits, and greatly reducing the number of large energy spikes. For simplicity, we shall denote the estimated decibel power spectrum

$$\hat{\Lambda}(f_k) = 10\log_{10}\left(\hat{S}(f_k)\right). \tag{2.40}$$

The use of a decibel magnitude scale gives spectra an appearance much closer to that of a normal distribution, and hence, this step should be taken if spectral moments are to be calculated. By using a decibel magnitude scale, another subtle consideration is revealed: where should the base-line, or "zero-reference" be positioned? That is, the moments should be calculated from the *normalised* spectral distribution $\Omega(f_k)$ where

$$\Omega(f_k) = \frac{\hat{\Lambda}(f_k) - r_z}{\sum_{n=1}^{N}\left(\hat{\Lambda}(f_n) - r_z\right)} \tag{2.41}$$

where $r_z < \hat{S}(f_k)$ for all $f_k \leq f_N$ is some arbitrary constant reference, and $2N$ is the window-size. Peculiarly, this does not usually seem to be mentioned in the literature.

There are two main considerations: firstly, the zero-reference must be sufficiently low that all spectra under consideration lie above it. It is unclear how to interpret regions where the spectrum drops below the zero reference. On the other hand, it is still desirable that the "tails" of the distribution approach the zero-reference near its edges. The first four moments will not characterise well distributions that have a high energy density at either end of the spectrum.

We shall see in Chapter 5 that simultaneously satisfying all the conditions in §2.3.3 and §2.3.4

is usually impossible, and that large concessions must be made in order to obtain a workable system. Additionally, basic measures to subtract known ambient-noise from speech recordings could be undertaken, but is rarely seen to be done in the literature.

### 2.3.5 A note on pre-emphasis

Occasionally in the literature, pre-emphasis of the signal is undertaken, with little explanation as to how or why.

Traditionally, pre-emphasis of the low-energy signal typically found at higher frequencies, was used to make better use of the amplitude resolution of the recording media. It is also used as a method of 'flattening' the spectrum prior to spectral estimation, in order to reduce the bias that can arise when estimating non-flat spectra.

In regard to the calculation of spectral moments, there may be a case for using pre-emphasis to optimise the spectral shape. As already mentioned, the closer to a nearly normal curve the spectral distribution under examination is, the more accurately the first four moments will describe the shape. Indeed, in much of the spectral moment literature, pre-emphasis is used prior to spectral moment calculation, but details are rarely given.

The most considerable problem facing a standard pre-emphasis step in spectral moment calculation is the potential diversity of fricative spectra. If a fixed pre-emphasis method is used, then we cannot hope to improve all spectra. If the methodology incorporates some nonlinear function of the spectral shape, then we cannot expect the spectral moments to be reliable.

### 2.3.6 Reconstruction using the Gram-Charlier expansion

In order to determine how well spectral moments have captured the shape of a distribution, it is first necessary to reconstruct a distribution from the moments and any other knowledge available from the 'results'. A comparison of the reconstruction to the original distribution will reveal how well (i.e. how *uniquely*) the spectral moments describe that distribution. It will indicate which features are well described, and where problems may arise.

Any distribution can be uniquely described by an infinite set of moments. Conversely, a finite set of moments can describe an infinite set of distributions. The Gram-Charlier expansion constructs the distribution of maximum likelihood for a given set of the first four moments (Kenney and Keeping 1964):

Assume the $x$-axis variable has been standardised, and denote it by $\phi = (x - \nu_1)/\sigma$.

By repeatedly differentiating the function $e^{-\phi^2/2}$ we obtain:

$$\begin{aligned}
\frac{d}{d\phi}(e^{-\phi^2/2}) &= -\phi e^{-\phi^2/2} \\
\frac{d^2}{d\phi^2}(e^{-\phi^2/2}) &= (\phi^2-1)e^{-\phi^2/2} \\
\frac{d^3}{d\phi^3}(e^{-\phi^2/2}) &= -(\phi^3-3\phi)e^{-\phi^2/2} \\
\frac{d^n}{d\phi^n}(e^{-\phi^2/2}) &= -(1)^n H_n(\phi)e^{-\phi^2/2}
\end{aligned}$$

where $H_n(\phi)$ is a polynomial in $\phi$, of degree $n$, called the $n$th Hermite polynomial. By repeated integration by parts, it is easy to show that

$$(2\pi)^{1/2}\int_{-\infty}^{\infty} H_m(\phi).H_n(\phi)e^{-\phi^2/2}d\phi = \begin{cases} n! & \text{for}\, m=n \\ 0 & \text{for}\, m \neq n \end{cases} \tag{2.42}$$

Hence if $\psi(\phi)$ stands for $(2\pi)^{1/2}e^{-\phi^2/2}$ and if we assume that a given frequency function can be expanded in a series

$$g(\phi) = c_0\psi(\phi) + c_1\psi'(\phi) + \cdots + c_n\psi^{(n)}(\phi) + \cdots \tag{2.43}$$

we can formally obtain the constants in the series by means of (2.42). Multiplying (2.43) by $H_n(\phi)$ and integrating term by term, we have

$$\int_{-\infty}^{\infty} g(\phi)H_n(\phi)dt = \sum_r c_r \int_{-\infty}^{\infty} \psi^{(r)}(\phi)H_n(\phi)d\phi = (-1)^n n! c_n \tag{2.44}$$

since all terms in the sum except for that which $r = n$ give zero on integration. Substituting $H_0 = 1, H_1 = \phi, H_2 = \phi^2 - 1, H_3 = \phi^3 - 3\phi, H_4 = \phi^4 - 6\phi^2 + 3$, we obtain

$$\begin{aligned}
c_0 &= \int_{-\infty}^{\infty} g(\phi)d\phi = 1 \\
c_1 &= -\int_{-\infty}^{\infty} \phi g(\phi)d\phi = 0 \\
c_2 &= \int_{-\infty}^{\infty} (\phi^2-1)g(\phi)d\phi = 0 \\
c_3 &= -\gamma_1/3! \\
c_4 &= (\alpha_4 - 6 + 3)/4! = \gamma_2/24
\end{aligned}$$

therefore

$$g(\phi) = \psi(\phi) - \frac{\gamma_1}{6}\psi^{(3)}(\phi) + \frac{\gamma_2}{24}\psi^{(4)}(\phi) - \cdots \tag{2.45}$$

This is the Gram-Charlier A Series. It has been shown that the series is not convergent except under rather restrictive conditions. However, the important point is that a few terms provide a good approximation to $g(\phi)$.

That is, the first few moments can be used to reconstruct a reasonable approximation to the distribution from which the moments were calculated. How well this reconstructed distribution matches the intended distribution will tell us how well the spectral moments can be expected to perform.

### 2.3.7 Spectral moments of large-variance spectral estimates

In §2.2, it was shown that spectra estimated from a single Fourier-transformed window of data will have a large error variance at any given frequency. If the estimated distribution has a large variance, then the higher-order moments can be expected to be less reliable, since they place increasing emphasis on variations in the tails of the distribution, where a large error variance is known to exist.

However, if the distribution under examination is not approximately Gaussian, moments calculated will become insensitive to changes in the distribution that we may wish to capture. The moments of a non-Gaussian distribution will be less sensitive to change anywhere in the distribution than if the same amount of change occurred in a distribution that was close to a Gaussian. For instance, an increase of 2dB in a specific region in a flat distribution will correspond to a much smaller change in the calculated moments than if the same increase occurred in the same region in a distribution that was close to Gaussian. These expected properties of spectral moments are demonstrated in Chapter 5.

## 2.4 Summary

Fricative analysis can be approached in a number of different ways. One approach is to construct mathematical models of the vocal tract. This relies upon detailed knowledge of noise sources and interactions. Turbulent noise sources however are highly complicated, and a limited amount is known about their behaviours, spectra, and interactions with other sources. With the introduction of obstacles into the path of turbulent jets, the complexity increases further. Mathematically modelling such a system becomes unmanageable.

Nonparametric analysis methods are a somewhat more appealing approach that have not been fully explored. The turbulence noise generated during fricative production should be treated as stochastic process, and yet often in the fricative analysis literature this is not performed suitably. Appropriate nonparametric spectral estimation methods are examined. A certain amount is known about fricative production, and analysis tools need to make use of all the information present, so it is important that the nonparametric data are as accurate as possible.

Well founded principles of good spectral estimation procedures for stochastic processes are often overlooked in the fricative analysis literature.

In order to obtain an unbiased estimate of time-series data, a suitable data taper must be used. In doing so, the amount of information being used to calculate the estimate is reduced, and this increases the variance of the error of the estimate. In the case of a modified periodogram estimate, the variance error of the estimate at any given frequency is greater than underlying quantity being estimated at that frequency. If fine spectral details are to be examined closely, this estimate is highly unsatisfactory.

In order to reduce the variance error, a *consistent* estimate is required: more data need to be incorporated into the calculation, without attempting to increase the amount of information within the estimate. While using a longer data-window increases the amount of data incorporated into the estimate, it also increases the resolution of the estimate, and since (without further treatment) this results in no overall reduction in the error variance, the modified periodogram is said to be 'inconsistent'.

Where the underlying system is known to be stationary, time-averaging can be used to generate a consistent estimate. Alternatively, ensemble-averaging methods can be used to reduce the variance of the estimate where the underlying system is known to be ergodic. Unfortunately, neither stationarity nor ergodicity can be verified unless the other is known to be true. In speech analysis, neither assumption would appear to be particularly likely however. If it is instead assumed that neither is true, alternative methods of obtaining a consistent estimate must be considered. In this case, frequency smoothing may be used; however, it can be shown that frequency smoothed estimates that reduce the variance by a significant amount, also introduce a significant amount of local bias.

Multitaper analysis provides an alternative method of improving the estimate. The amount of data incorporated into the estimate is maximised, while the local bias is minimised. Multitaper analysis excels over the other methods where spectral estimation over short time intervals of non-stationary non-ergodic processes is required. It therefore seems likely to be very well suited to fricative analysis.

Spectral moments have previously been implemented in a number of different ways. Frequency range selection and magnitude scale selection are two of the most obvious choices that are likely to be significant in determining the effectiveness of the spectral moments. While various parameters have been used in previous studies, no attempts appear to have been made to investigate the effects these parameters have on the performance of spectral moments.

The Gram-Charlier expansion allows the distribution of maximum likelihood to be constructed given the first four moments. The first four moments are an incomplete basis, and hence any given set of the first four moments will describe all of an infinite number of different distributions. The Gram-Charlier expansion can be used to clarify which features of a distribution have the greatest levels of influence over the spectral moments that are calculated.

# Chapter 3

# Method

The speech of cochlear implant users is the subject of much examination. In particular, studying the effects on fricative production of partially-known changes to the auditory system by cochlear implant device provides us with information about the production and perception systems. However, tools for describing and analysing fricative productions are of limited availability. Such tools would ideally be able to measure changes and differences of both normal and disordered fricative productions.

In order to continue, some sample speech data from subjects with normal hearing, and from subjects with abnormal hearing is required. We begin by considering what criteria the subjects should satisfy.

Since analysis of the disordered speech of cochlear implant users is a significant incentive for this work, we begin by presenting information about cochlear implant subjects whose speech and hearing backgrounds are well documented. In order to evaluate whether new fricative production measurements are suitable for use with disordered speech, only a limited number of tokens are required. This coincides with the amount of speech data readily available from cochlear implant subjects: intensive speech recording sessions of speakers with some hearing or speech impairment is often more stressful than for subjects of normal speech and hearing.

The criteria by which subjects of normal hearing are chosen may be based upon the backgrounds of the cochlear implant subjects, amongst other considerations. The speech of the normal hearing subjects is to be analysed in order to discover typical variabilities across fricative tokens; a suitable corpus is discussed.

Finally, the methods for recording, storing and editing speech data are presented. Equipment descriptions, procedures used during recording sessions, and post-recording data processing such as data segmentation and calibration, are included.

## 3.1 Subject Requirements

Being able to measure changes in the speech of subjects whose production is affected by a known cause of hearing-loss would be most useful. For example, cochlear implant subjects offer an excellent opportunity for understanding the relationships between hearing and speech. To commence analysis of these relationships, it is necessary to acquire one or more subjects exhibiting interesting speech disorders. A small number of subjects of varying speech development backgrounds may be selected in order to explore different kinds of production variations.

To describe the degree of 'abnormality' of disordered speech, or to measure changes that may occur after some known change in hearing status (cochlear implant activation, for example), it is first necessary to establish a measure of 'expected values' from an analysis of normal speech. Therefore, 'normal' speakers are required so that a reasonable sample of 'normal' fricative productions can be obtained. A sufficient number of these subjects should be present in order to gain a coarse measure of typical ranges and variabilities.

### 3.1.1 Cochlear implant subjects

In order to improve the performance of a cochlear implant, users undergo routine checkups. The subject's hearing and speech perception is evaluated, and parameters of the device are often adjusted. An appointment for hearing evaluation is also usually undertaken prior to implant insertion. These routine evaluations of hearing and speech perception are also a convenient time to attempt to measure any changes in the speech production of subjects.

Subjects taking part in routine cochlear implant adjustment procedures at the University of Southampton Cochlear Implant Centre are readily available test subjects with well-documented hearing disorders. Of these, two males (MCI-13 and MCI-14) and two females (FCI-15 and FCI-16) with cochlear implants are used for generating test-data for later measurement. These subjects were chosen for their different speech and hearing backgrounds, since this should provide us with data that can be used to test various production description methods.

#### 3.1.1.1 Subject selection

All subjects are English speakers who were classified as having profound deafness prior to implantation. Unless otherwise specified, subjects' deafness was post-lingual. All subjects were implanted with Cochlear CIN-24 internal electrode arrays. Specifics about age of implantation and processing strategies for the individual subjects follow.

Subject MCI-13 was implanted at 66 years of age, using an ESPRIT-24 processor device implementing the SPEAK processing strategy. Speech data for this subject were taken from a recording one year post implantation.

Subject MCI-14 received a cochlear implant in his left ear at age 66, implementing the SPEAK strategy on an ESPRIT-24 processor. This subject does not originate from the South of England, and can be considered to have a slight Northern-regional accent. The data from recordings made one year post implantation are used for this subject.

Subject FCI-15 lives in the South of England. At age 35 she was fitted with a Nucleus CI24M in her left ear; (total insertion of the electrode array was achieved). During her initial tuning week, she tried the CIS and ACE strategies of her ESPRIT-24 speech processor, and received speech and language therapy for approximately 10 months, after which she changed to using the

SPEAK speech processing strategy. Data used here are from recordings made at one and two year post-implantation.

Subject FCI-16 was diagnosed with hearing loss due to neonatal jaundice at 3 years of age, and hence deafness may be considered pre-lingual. Progressive hearing loss continued until profound deafness became established at 17 years of age. At this time the subject was fitted with an ESPRIT-24 processor (SPEAK strategy). This subject has predominantly lived in the South of England. Data are taken from recordings made one year post implantation.

#### 3.1.1.2 Corpus

Only real words are considered. This effectively eliminates an extra degree of variability in the interpretation of 'nonsense' words by speakers, which are not of interest, and may mask other variabilities. In particular, from past experience it has been found that subjects with affected hearing often have much greater difficulty than normal-hearing subjects when faced with the task of reading a rhyming list of mixed real and nonsense words. It seems that while normal-hearing subjects are immediately able to see the pattern of similarity in the expected sound for the list of words, this does not always occur in the speakers with affected hearing. In order for the results of normal speakers to be comparable with the results of speakers with affected-hearing, a real word corpus seemed to be the best choice.

The corpus used to evaluate the speech of cochlear implant subjects for this, and other studies, consisted of the following sections:

- 15 sentences of "It's a h/V/d again.", where /V/ ∈ {/i, ɪ, ɛ, æ, ɑ, ɔ, u, ʊ, ʌ, œ, o, e, ɑj, ɑw, ju/}.
- 6 sentences of "It's a /C/od again.", where /C/ ∈ {/p, t, k, b, d, g/}.
- 15 short sentences.
- The Rainbow passage (Fairbanks 1960).
- 14 lists of common words, each containing approximately 16 words (Parker 1999).
- The Dog and Duck passage (Parker 1999).
- In some cases, a further passage titled "Sue's Seaside Trip", which focusses on the subject's production of all sibilants, using many words from Parker's lists.

This corpus was designed to capture a range of different speech characteristics in disordered speakers. Productions of only a few words from the Parker (1999) word lists were used for the analysis in Chapter 7.

### 3.1.2 Normal hearing subjects

We wish to estimate typical fricative production variability within and across vowel contexts for a number of tokens and speakers. It was decided to limit variability initially by using subjects of same gender, age range and accent background.

#### 3.1.2.1 Subject selection

Seven male subjects (M-00 through to M-06), and six female subjects (F-07 through to F-12) between the ages of 20 and 30 were chosen who all lived in the south of England since birth and have native British English-speaking parents. One of the male subjects (M-00) whispered most of the corpus, and so all data for this subject were discarded. Only one subject (M-03) had any special phonetic knowledge.

The limitations on regional accent were imposed for several reasons. Firstly, since the effects of vowel context on fricative production is under examination, some control on the vowel context is needed, and this would not be possible if regional accents were included. Additionally, the cochlear implant subjects have similar regional accent backgrounds, and so this limitation provides us with good comparative material.

#### 3.1.2.2 Corpus

A sample of speech including fricatives is required in order to allow us to observe the behaviour of existing characterisation metrics, demonstrate the properties of spectral estimation techniques when used with fricative signals, and develop new methods of characterisation.

The corpus is designed so that the English voiceless fricatives and voiced fricatives can be studied in real words containing /$V_1$F $V_2$/ contexts where /F/ ∈ {/f,θ,s,ʃ,v,ð,z,ʒ/}, /$V_1$/,/$V_2$/ ∈ {/i,u,ə/}. Symmetrical contexts (where $V_1$=$V_2$) could only be found for all fricatives when $V_1$=$V_2$=/i/. The set of vowel contexts used thus consists of /iFi/, /iFə/, /uFi/, /uFə/, /əFi/ and /əFu/. Only real words were included in the corpus, for the purposes described in §3.1.1.2. This set of vowel contexts is being referred to when 'all vowel contexts' is specified later in the text. Tables of words in which the desired /VFV/ combinations appear, are given in Appendix A.

The context has been fixed as /$V_1$F $V_2$/, largely to keep the problem of segmentation consistent. While it is recognised that fricatives often occur in clusters and other contexts, it is not necessary to tackle all the issues of segmentation and coarticulation in order to begin work on better fricative analysis methods.

In order to simulate typical (but not extreme) variations in production, the order of words was varied. Each page was designed so that a given word will appear both at the start, at the end, and at every point within a line of words to be read (see Appendix A). The first page of words was used as a test-page, to familiarise the subjects with the words to come, and also to allow the recording gain to be adjusted (see §3.2).

## 3.2 Data acquisition

The following methods of speech recording and analysis were consistently applied for all seventeen subjects.

### 3.2.1 Recording equipment and setup

Recordings were made in a sound-proofed quiet-room. The subject was seated in a chair with headrest, with a music-stand placed in front of them displaying the corpus. A Brüel & Kjær (B&K) 4133 microphone, fitted to a B&K 2639 preamplifier, was held by a floor-stand at a distance of 1m from (and directed towards) the subject's mouth. The output of the preamplifier was fed into a B&K 2636 amplifier, which was set to give a bandpass of 22 Hz to 22 kHz. The output of the amplifier was connected to one channel of a Sony DAT TCD-D7 corder with a sampling rate of 48 kHz, with 16-bit amplitude resolution. The recording setup is shown in Figure 3.1. The music stand was positioned underneath the microphone, so as to minimise interference with recording. A laryngograph was also used during recording sessions of cochlear implant subjects. This signal was fed into a separate channel of the DAT, but the laryngograph data are not used in this work.

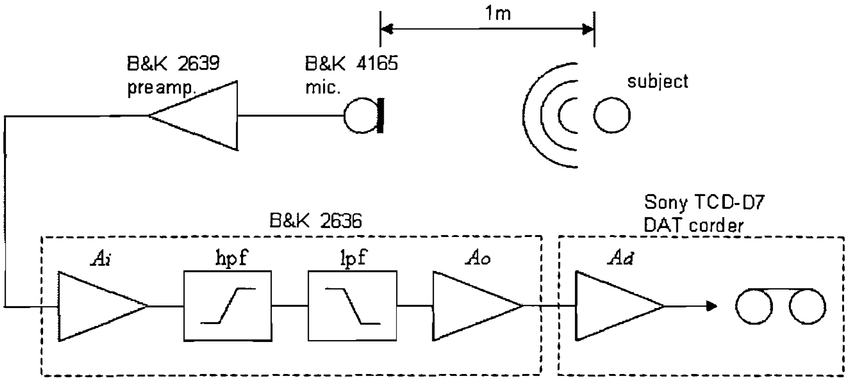


FIGURE 3.1: Recording Setup.

For each subject, the B&K 2636 amplifier's input and output gains ($A_i$ and $A_o$ respectively) were both initially set to 20dB. While the subject read the first test-page of corpus, the gain of the TCD-D7 was adjusted so that levels did not quite reach maximum limits, where the signal level would become clipped. If the speaker was particularly loud, and the minimum gain on the DAT was not low enough to prevent signal clipping, the B&K 2636 gain settings were reduced by 10dB. The subject would then read the entire corpus, page by page. Subjects were instructed to sit still, and attempt to keep their head in the same position throughout the session.

After the corpus had been read, 30s of ambient noise were recorded. Next, $A_i$ and $A_o$ were reduced in (calibrated) levels of 10dB so that the 60s of calibrated 94dB SPL test-tone signal (produced by B&K 4230 calibrator) was presented to the TCD-D7 at a suitable level. The test-tone was recorded so that all signals could be calibrated to absolute SPL at a later stage. Settings of all equipment were noted for each subject to aid calibration procedures.

### 3.2.2 Data storage and initial segmentation

All data were transmitted to PC hard-disk from the DAT, and stored as WAV-format files. Each token spoken during the session for each subject was manually edited using Syntrillium's 'Cool Edit 2000' to a small file containing the fricative and about half of the preceding vowel and of

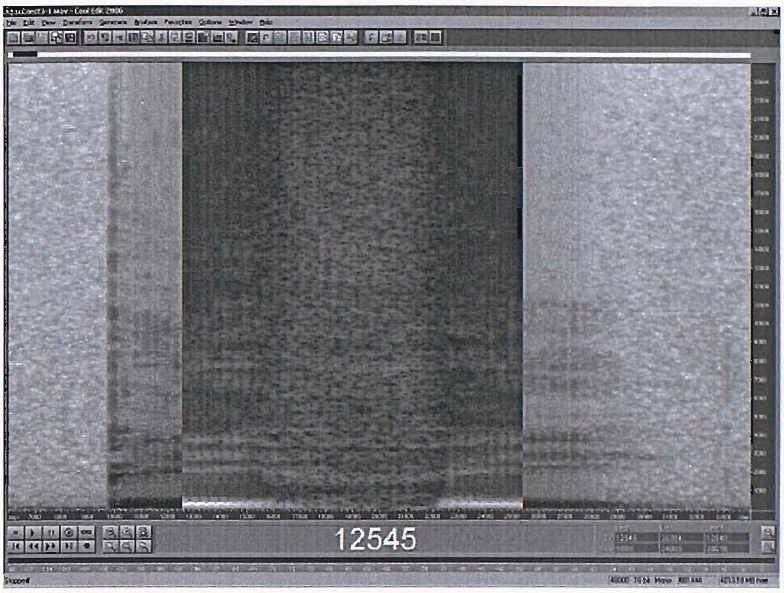


FIGURE 3.2: Example of selection of /ifi/ fricative segment from the word 'beefy' using Cool Edit 2000. The dark, 'inverse' central block is the selected segment.

the following vowel. This was done to ensure the entire fricative would certainly be captured, and so that transients occurring at the fricative boundaries could later be studied. It would also allow some automatic segmentation algorithm to systematically locate a 'central' region of each fricative (by some arbitrary definition), which would allow large volume processing to be performed automatically.

The 'spectral view' was used to determine the approximate location of the centre of the vowels surrounding the fricative, and save these short /VFV/ segments of data, as indicated in Figure 3.2. If the vowels were of differing lengths, less of the longer vowel would be included, so that the fricative remained in the centre of the segment. This procedure was performed for each of the 1,728 voiceless fricative tokens by all male and female normal-hearing speakers, and also for some sample tokens from the speakers with cochlear implants.

### 3.2.3 Data alignment

An automated system was required to capture, and save, a 'central' region of each voiceless fricative of the 1,728 /VFV/ segments of the normal-hearing subjects. These 'central' regions of the fricative can be referred to as fricative 'tokens', and are much more straightforward to work with when large numbers of tokens are being processed.

We define the boundaries of each fricative token as a point where the frication noise becomes sufficient, compared to the frication occurring in the vowel segments on either side. A suitable measure of frication noise is straightforward to acquire using the method described by Scully, Castelli, Brearley, and Shirt (1992), whereby the signal below 3.9 kHz is filtered out, and the magnitude of the remaining signal is the measure of frication noise.

The frication noise over time $\Phi(n)$ for each /iFi/ production segment was calculated over $N$ 512-point (10.6-ms) data windows, advancing by 256 points (5.3 ms) through each /VFV/ segment.

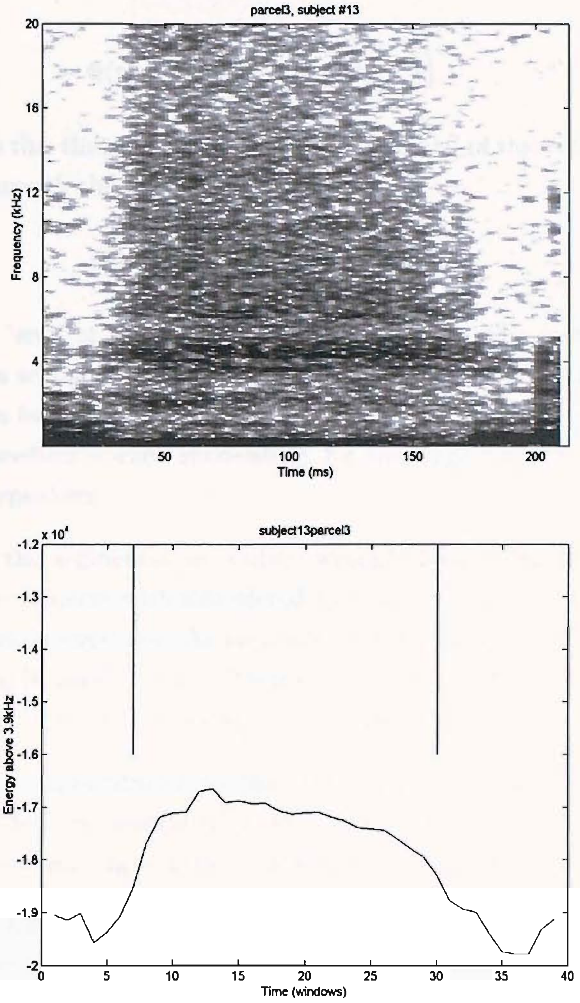


FIGURE 3.3: Example /ɑsə/ segment from the word 'parcel' produced by subject MCI-13. Top plot shows spectrogram of segment. Bottom plot shows total 'frication noise'. Vertical lines are explained in the text.

A typical plot of $\Phi(n)$ through an /ɑsə/ segment from the word 'parcel' produced by subject MCI-13 is shown in Figure 3.3. The low levels of frication noise at the edges indicate the vowel regions, and the rise in frication noise is straightforward to see.

The value of maximum frication $\Phi_{(\max)}$ in the central 30% of each file,

$$\Phi_{(\max)}(n_{(\max)}) = \max_{N/3<n<2N/3} \{\Phi(n)\} \tag{3.1}$$

was located, where $n_{(\max)}$ is the value of $n$ at which this maximum frication occurs. The minimum values of frication either side of this, were calculated

$$\Phi_{(\text{lmin})} = \min_{n<n_{(\max)}} \{\Phi(n)\}\,, \tag{3.2}$$

$$\Phi_{(\text{rmin})} = \min_{n>n_{(\max)}} \{\Phi(n)\}\,. \tag{3.3}$$

The function $\Phi(n)$ was stepped through systematically for $1 < n < N$. The data for $x(n)$ were discarded until

$$\Phi(n) > 0.5\left(\Phi^{(\mathrm{max})} + \Phi^{(\mathrm{lmin})}\right). \tag{3.4}$$

At the value of $n$ where this threshold was exceeded, the 'start' of the fricative token was defined. We continue to step through the data until

$$\Phi(n) < 0.5\left(\Phi^{(\mathrm{max})} + \Phi^{(\mathrm{rmin})}\right) \tag{3.5}$$

at this value of $n$ the 'end' of the fricative token is defined. The vertical bars in Figure 3.3 indicate where the data segmentation has taken place relative to the frication levels over time of this production. As can be seen, the segmentation procedure works well in this case. In fact, this data segmentation procedure worked successfully for the large majority of fricative production of the normal hearing speakers.

Occasionally however, the segmentation routine wrongly locates the start and end points. In particular, data for the subjects with disordered speech often produce tokens that are not well suited to this segmentation approach. As an example, consider subject FCI-15, who often forms complete closure during [s] productions. The frication levels, and automatic segmentation start and end points for such a production are shown in Figure 3.4.

Results of the automatic segmentation routine were scrutinised, and on the occasions where it had unsuccessfully located the central fricative region satisfactorily, the start and end points were defined by hand, attempting to retain the overall segmentation criteria.

The resulting fricative tokens are therefore of a variable length, as expected. In addition, this segmentation procedure tends to include some of the transition regions, which appear to be important in the identification of at least some fricatives. In some later chapters, more central regions of the fricative tokens are required for analysis. In these cases, the central portions of these fricative tokens are very straightforward to locate automatically and use.

### 3.2.4 Calibration and filtering procedures

The process for converting the recorded data for each subject into standard units of dB SPL was largely automated using MatLab 6. Files containing the recording of the 94dB SPL test-tone (we denote with variable '`tt`'), the ambient noise recording ('n'), and the speech recording ('x') were loaded. The total B&K 2636 gain set during speech and ambient noise recording ($Ai$[s] + $Ao$[s]), and during the calibration test-tone recording ($Ai$[c] + $Ao$[c]) were noted for every recording session. The difference ($Ai$[s] + $Ao$[s] − ($Ai$[c] + $Ao$[c])), was then stored ('`tt_deficit`'). The following operation then took place each time speech data from a specific recording session were analysed:

```
%Load relevant files
[n,nfs,nbits]=wavread('amb_noise.wav');
[tt,ttfs,ttbits]=wavread('testtone.wav',win_len);
tt_deficit=load('tt_deficit.mat');
```

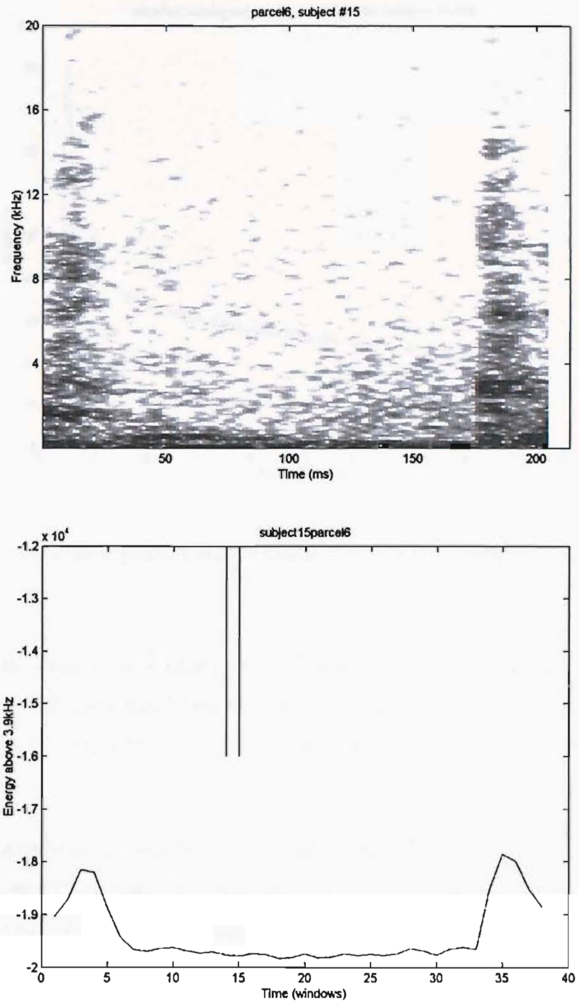


FIGURE 3.4: Example /ɑsə/ segment from the word 'parcel' produced by subject FCI-15. Top plot shows spectrogram of segment. Bottom plot shows total 'frication noise'. Complete closure during the [s] production leads to the segment location method being unable to cope.

```
%Calibration procedure...
tt=tt*(10^(tt_deficit/20)); % Convert tt_deficit value from dB
tt_energy=fft(tt,win_len);

%Calculate energy in test tone
tt_psd=2*(abs(tt_energy(1:win_len/2)).^2)/(win_len*ttfs);
tt_power=sum(tt_psd);
cal_boost=((10^(94/10))/tt_power)^(1/2); %Boost factor

n=n*cal_boost;
x=x*cal_boost;
test=tt*cal_boost;
```

The spectrum of the newly 'calibrated' test tone could be observed from 'test', so that the expected response to a 94dB signal could be checked. The response calculated using a 512-point

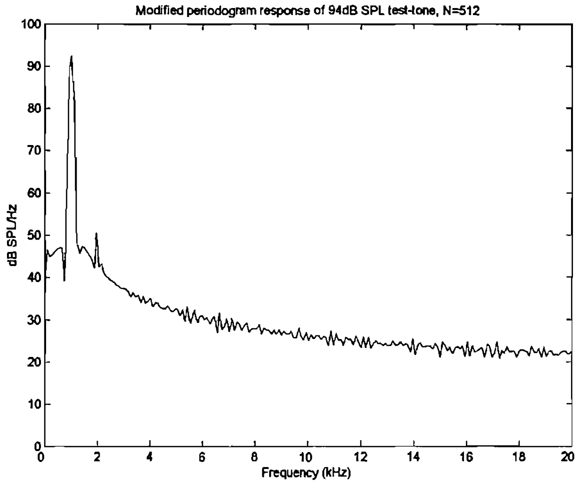


FIGURE 3.5: Modified periodogram response of 94dB SPL test-tone, $N$=512.

modified periodogram is shown in Figure 3.5. Note that the peak does not quite reach 94dB SPL, due to leakage. Multitaper analysis would of course not usually be a suitable method for measuring the spectrum of such a tone, but in this case it serves as a guide for analysis performed later.

Filtering was also automatically performed using MatLab 6. In order to attenuate unwanted low-frequency room noise (that can be seen on spectral plots in Chapter 4) and energy above 20 kHz (resulting from the 48-kHz sampling frequency of the DAT), frequency bins outside the range of interest were discarded. Unless otherwise indicated, all data were filtered to remove information below 216 Hz and above 20,063 Hz.

# Chapter 4

# Analysis: spectral estimation of fricatives

The properties of each of the spectral estimation methods described in §2.2 are now explored from a practical perspective. Recorded acoustical speech data of subjects M-01–M-06 are used for this purpose, so that only normal characteristics of each spectral estimation method will be present.

We examine the assumptions upon which the various averaging techniques discussed in Chapter 2 are based. The problems encountered when these assumptions break down are demonstrated, and commonly seen features in spectral estimates and spectrograms are shown to often be artifacts of the estimation technique.

Finally, recommendations for the spectral analysis of fricatives are discussed, based on the observations presented.

## 4.1 The first estimate

Figure 4.1 is the power spectral estimate calculated from a 512-point rectangular window placed in approximately the centre of the fricative /s/ in the word "fleecy" spoken by subject M-04.

As shown in §2.2.1.3, this estimate of the spectrum is biased and has a large variance. The bias is due to the large side-lobe leakage that occurs when using a rectangular window, so that the relatively high energy densities in the lower frequency regions of the signal leak into neighbouring regions. This gives the impression of greater energy than actually exists in regions where the energy content is actually low.

### 4.1.1 Reducing side-lobe leakage: the modified periodogram

This bias is reduced significantly by multiplying the original data by a window that has a response with attenuated side-lobes. A commonly used family of windows are Blackman-Tukey windows

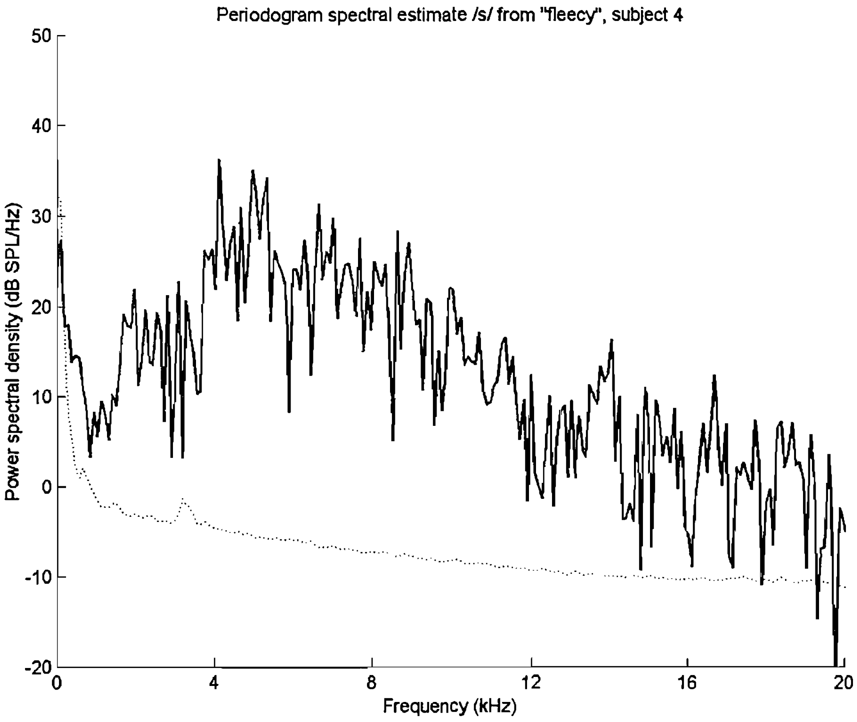


FIGURE 4.1: Periodogram spectral estimate using using 10.6-ms (512-point) Daniell window from centre of [s] from production of "fleecy" by subject M-04. Dotted line indicates ambient noise.

$\lambda(k) = 1 - 2r + 2r\cos(\pi k/M)$ for $|k| \leq M$, and 0 otherwise, where setting $r = 0.25$ results in the *Hanning* window, and $r = 0.23$ the *Hamming* window. Figure 4.2 shows the power spectral density estimate of the same data windowed with the Hanning window, superimposed onto the previous estimate. Notice that the energy trough in the 100-Hz–1.5-kHz frequency region has dropped significantly in energy (presumably towards the underlying values). Originally these would have been biased by the leakage from the large-valued peak at 2 kHz.

This new estimate, having improved bias properties, is commonly used in speech analysis (e.g. O'Shaughnessy 1987). Yet as mentioned in §2.2.1.3, if the signal under analysis is a fricative, then the process must be considered stochastic, and so the variance of this estimate is still large. This variance can be demonstrated by superimposing a 95% confidence interval on a periodogram spectral estimate. Figure 4.3 shows a modified spectrogram of a 512-point section from the centre of [f] in a production by subject M2 of the word "beefy", with 95% confidence bounds. Considering the possible values from one frequency to the next, there are often large overlaps in the frequency response, so it is quite reasonable to assume that, where $d\hat{G}/df$ is positive at some $f = f_h$, $dG/df$ may in fact be negative at this frequency. This combines with the effect near peaks of very large variance in the estimate, to make locating maxima and minima from this estimate subject to large error. This in turn means that searching of peaks and troughs is error prone. Since a common aim in speech research is to locate peaks that may result from one or more poles in the vocal tract, and in the case of fricatives, the location of troughs resulting from zeros caused by antiresonances, an estimate with large variance is of questionable usefulness in these areas.

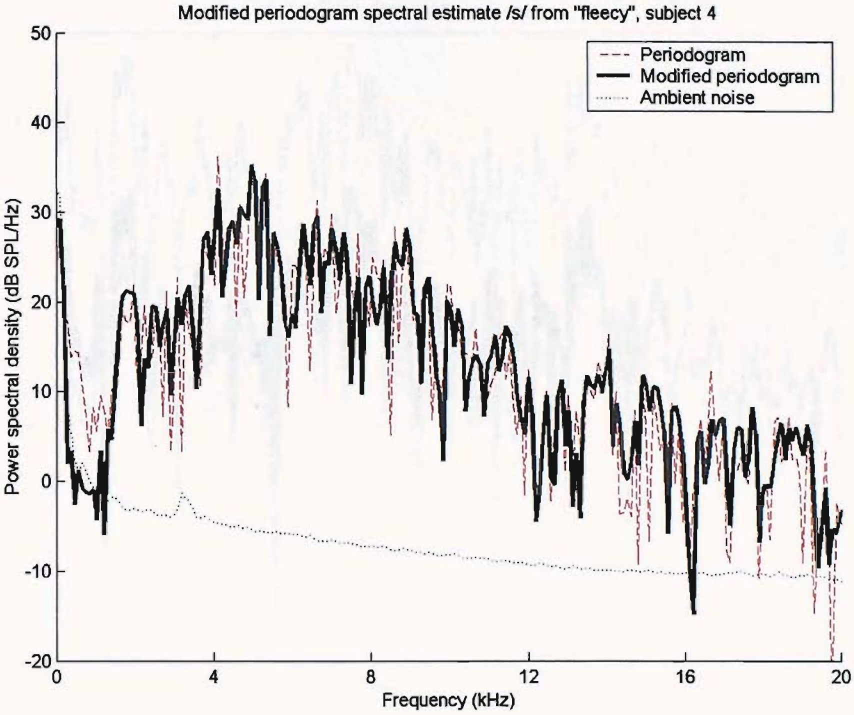


FIGURE 4.2: Modified periodogram spectral estimate using 10.6-ms (512-point) Hanning window, superimposed on unmodified periodogram (dashed). Taken from centre of [s] from production of "fleecy" by subject M-04. Dotted line indicates ambient noise.

Another important aspect of speech is the way in which the signal properties change over time, due to the constantly changing nature of the production system. Indeed, a common theory is that much of the information in speech is contained within the change in signal rather than the signal itself. It is therefore useful to be able to track changes over time of the underlying production mechanism from changes in the resulting acoustical signal properties over time. Commonly the peaks in speech spectra can be tracked, and the orientation of the production mechanism determines the position of these. It should also be possible to track the changes in orientation of the production mechanism during fricative production from the signal spectra so that, for example, it can be determined at which points in the production of the fricative the mechanisms are stationary, and at which points they are subject to more rapid change. In order to do this, the spectrum is computed from one time interval to the next. However, due to the large variance associated with this first estimate, changes in the spectrum that are comparable in size to the (considerable) variance of the estimate are impossible to recognise: the variance of the estimate overwhelms changes in the underlying system spectrum.

### 4.1.2 Estimate variance: white dots in the spectrogram

Again referring to Figure 4.3, note that some estimates of the spectrum that have a large negative deviation appear as particularly deep spikes up to about 20dB in size. There seem to be no similar positive spikes however. From the equations governing the estimate variance, both large positive and large negative errors may be expected.

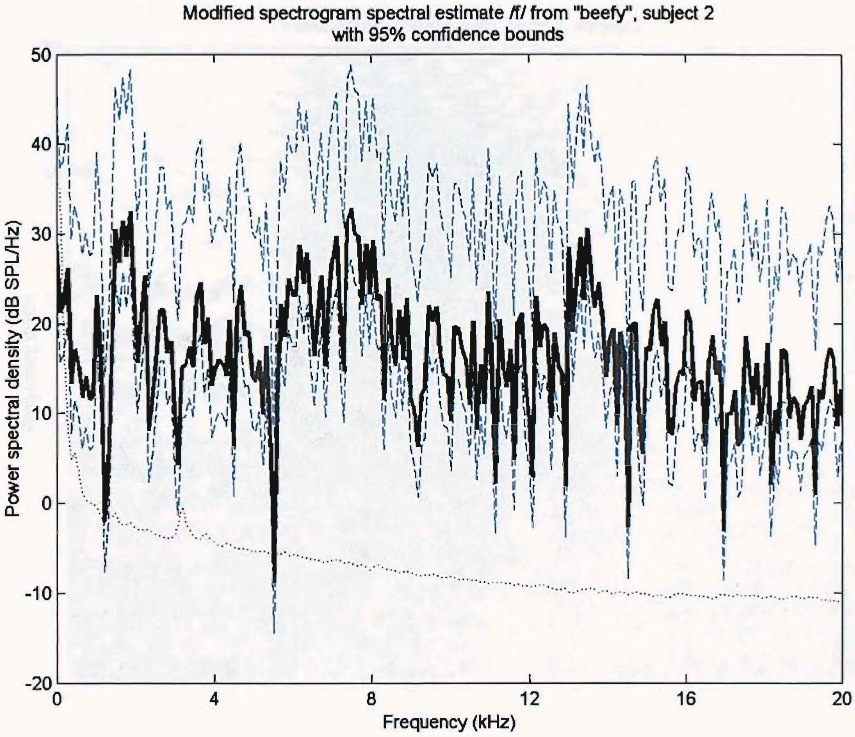


FIGURE 4.3: Modified periodogram spectral estimate with 95% confidence bounds. 10.6-ms Hanning-windowed section taken from centre of [f] from production of "beefy" by subject M-02.

In fact, the presence of these severe negative spikes is a result of the common practice of using a logarithmic scale when displaying power spectra. They give the spectral estimate a certain 'asymmetrical spiky' appearance. These spikes show up as undesirable artifacts in generated spectrograms.

Spectrograms are formed by calculating spectral estimates at incremental periods over an acoustical signal. In speech the spectrogram has become a powerful tool that allows a good visual representation of formants and speech dynamics. During fricative segments, the spectral estimates that are used to form the spectrogram are subject to increased variance, and as a result, the aforementioned large negative spikes appear. A common practice to increase the clarity of spectrograms is to define some baseline power value that will be set as white, use the maximum power value to define black, and use linear grey scale between these values. The baseline is usually set high enough that it intersects with the large negative spikes (as it usually is), so that the resulting spectrogram is peppered with white dots, as demonstrated in the typical spectrogram shown in Figure 4.4.

In order to resolve these issues, some method of reducing the variance of estimates is required.

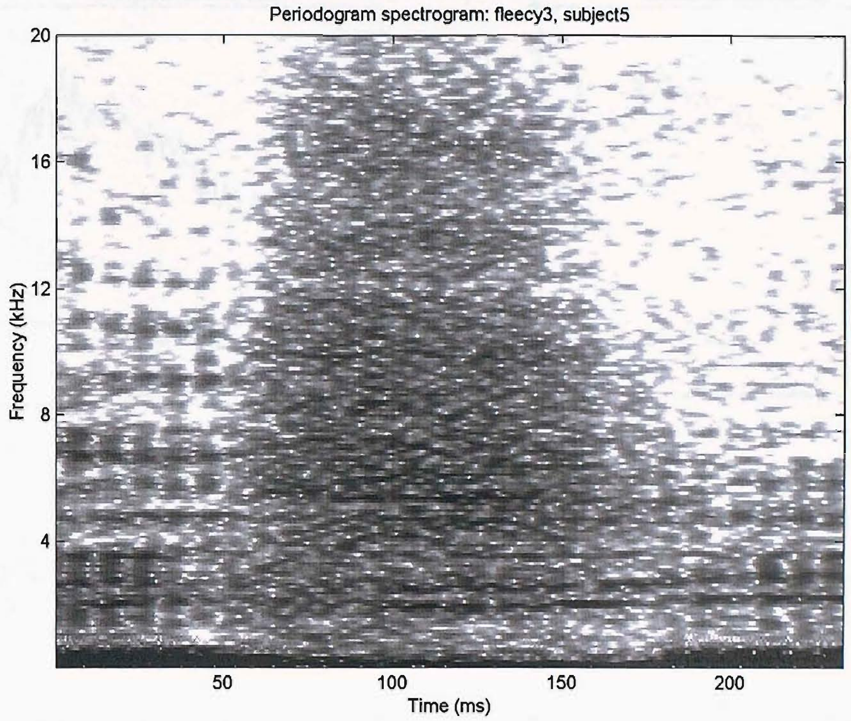


FIGURE 4.4: Spectrogram of [isi] production from 'fleecy' by subject M5. $N = 512$, overlap= 480. Notice the appearance of white dots in the fricative portion.

## 4.2 Time-averaging and fricative stationarity

As mentioned in §2.2.2.2, a common method used to reduce the estimate variance of a stationary process is to average the spectra calculated at several different times during a stationary portion of the stochastic signal. In speech analysis, this process presents several limitations.

Firstly, it requires that some portion of the fricative of interest is stationary. If it was known *a priori* that this assumption holds, then a much improved estimate could be assured. However, in trying to ascertain which parts of the fricative are stationary, and which are not, a spectral estimate with small variance of each segment is required. A time-averaged estimate formed over a non-stationary region of the process will give spurious results, from which little can be deduced with assurance.

What may be expected in a time-averaged spectral estimate formed over non-stationary data? Consider the position of a distinctive spectral peak. If this peak maintains amplitude but changes frequency smoothly over the course of the time-averaging period, this would be represented as a single broad energy band at somewhat lower magnitude. Indeed, it may not be recognisable as a peak at all. Alternatively, if the peak 'jumps' from one frequency to another, this will be represented as a double-peak. Of course, any actual broad energy band in the underlying system will appear in the estimate as a broad energy band, while any actual double-peak will also produce a double-peak in the spectrum. These qualities of the time-averaged spectral estimate make it particularly difficult to interpret over sections of data that are suspected to contain non-stationary components, or equally over data that are of unknown stationarity.

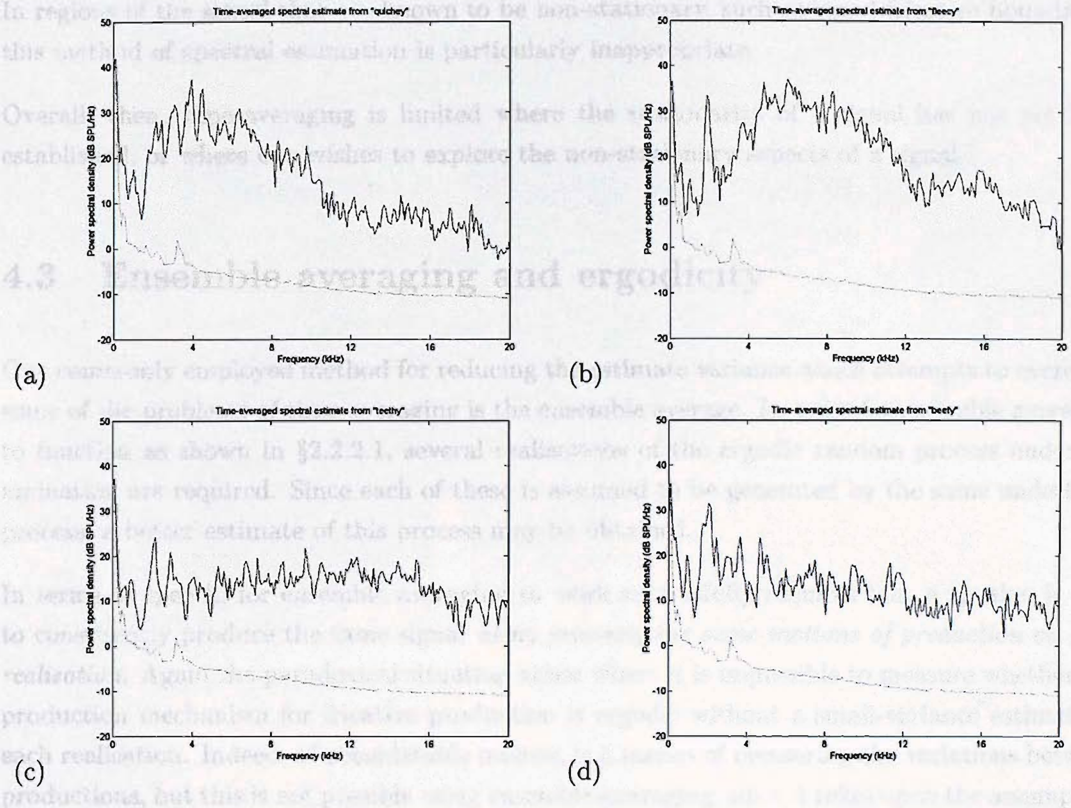


FIGURE 4.5: Time-averaged spectral estimates using 6 non-overlapping 512-point (10.6-ms) windows positioned approximately mid-fricative. (a) [ʃ] from centre of "quichey". (b) [s] from centre of "fleecy". (c) [θ] from centre of "teethy". (d) [f] from centre of "beefy" produced by M-01.

The spectral estimates shown in Figure 4.5 serve to highlight some of the potential uncertainties that are faced when interpreting the results of time-averaging. The large number of peaks that appear must now be interpreted as positions of the peaks over the averaging interval. As an example, consider the time-averaged spectral estimate shown in Figure 4.5.a, which has been generated by averaging 6 adjacent Hanning windowed segments of data in the middle of the fricative /ʃ/ in the word 'quichey', produced by subject M-01. The first clue that this estimate is unusual is the appearance of a double-peak at approximately 2.2 kHz. We may postulate that this double peak is in fact the time-averaged representation of a single peak that has moved over the time-averaging interval. Conversely, the strong peak around 2 kHz in the spectral estimate of /s/ from 'fleecy' in Figure 4.5.b, strongly suggests the position of a resonant frequency, although any amplitude changes that occurred have been averaged. Similar occurrences can be seen in the last two spectra of Figure 4.5. In §4.5.1 it is demonstrated that the 'ghost' peak just below the main peak at 2 kHz in Figure 4.5.d is actually due to a spectral peak that increases in frequency and amplitude during the course of the time-average.

The central region of the fricative is usually selected for calculating a time-averaged estimate, since it has been supposed that this is the region of greatest stationarity. However, the degree of stationarity in this region must be evaluated thoroughly. For the task of determining which parts of the fricative are stationary, some other method of minimising the estimate variance must be used.

In regions of the signal that are known to be non-stationary, such as vowel-fricative boundaries, this method of spectral estimation is particularly inappropriate.

Overall, then, time-averaging is limited where the stationarity of a signal has not yet been established, or where one wishes to explore the non-stationary aspects of a signal.

## 4.3 Ensemble averaging and ergodicity

One commonly employed method for reducing the estimate variance which attempts to overcome some of the problems of time-averaging is the ensemble average. In order for ensemble averaging to function as shown in §2.2.2.1, several realisations of the ergodic random process under examination are required. Since each of these is assumed to be generated by the same underlying process, a better estimate of this process may be obtained.

In terms of speech, for ensemble averaging to work successfully requires that a speaker is able to consistently produce the same signal *using precisely the same motions of production on each realisation.* Again the paradoxical situation arises where it is impossible to measure whether the production mechanism for fricative production is ergodic without a small-variance estimate of each realisation. Indeed, of considerable interest is a means of measuring the variations between productions, but this is not possible using ensemble-averaging, since it relies upon the assumption that all productions are identically produced.

Another difficulty that arises in forming ensemble-averaged estimates is that of identifying equivalent events in two separate fricative productions. Unfortunately, the start and end points of a fricative are ill-defined. Any definition for these points must be anchored to some known specific event in the production mechanism for the process.

Figure 4.6 shows time-plots of six productions of /iʃi/ from the word ‘quichey’ by subject M-06, on identical time and amplitude scales. It can be seen by eye that these fricative productions are of varying duration. This is not a rigourous analysis. However, the fact remains: no matter what definition of duration is defined, each of the above productions will be of different length. This at once suggests that the production mechanism — at least for this speaker producing this fricative — is non-ergodic.

A typical procedure for generating an ensemble average is demonstrated by Shadle, Moulinier, Dobelke, and Scully (1992). ‘Start’ and ‘end’ event labels are defined using the vowel-fricative transition, and fricative-vowel transition. Other methods incorporate use of an electromyogram (EMG) across the larynx to more accurately determine voicing onset and offset positions. Since the time intervals between these event labels are subject to some variation, a system of ‘temporal warping’ is needed to make the durations uniform, so that ‘events’ across productions can be located. Of course, such ‘temporal warping’ must be rather an arbitrary stage, and so must be treated as liable to produce misinformation.

An ensemble-averaged spectral estimate formed from central fricative portions of [ʃ] from six productions of ‘quichey’ by subject M-01 is shown in Figure 4.7, along with the ensemble-averaged spectra for three other unvoiced fricatives by this subject.

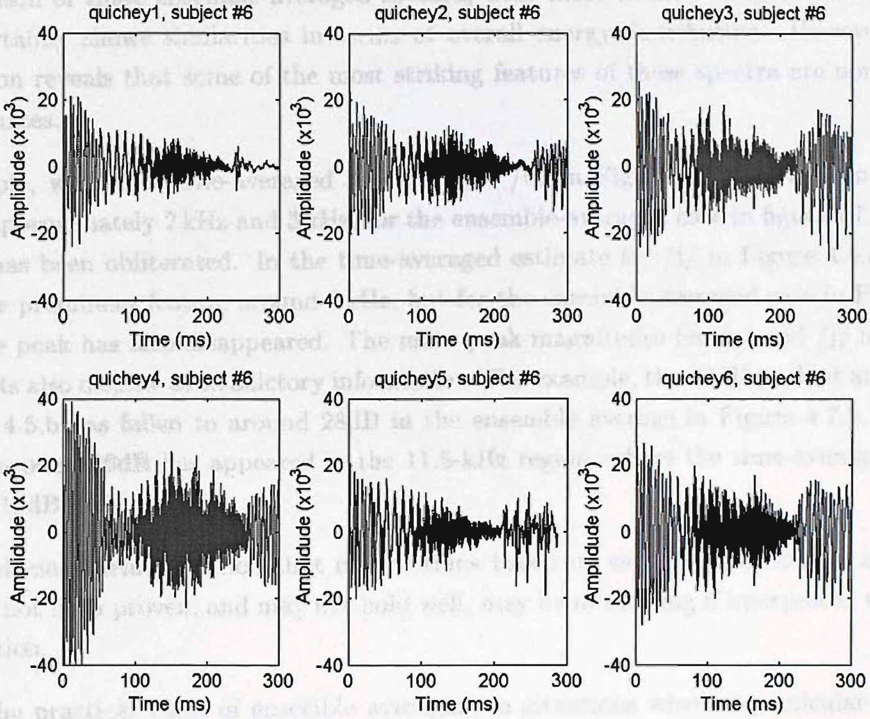


FIGURE 4.6: /iʃi/ from six "quichey" productions, subject M-06.

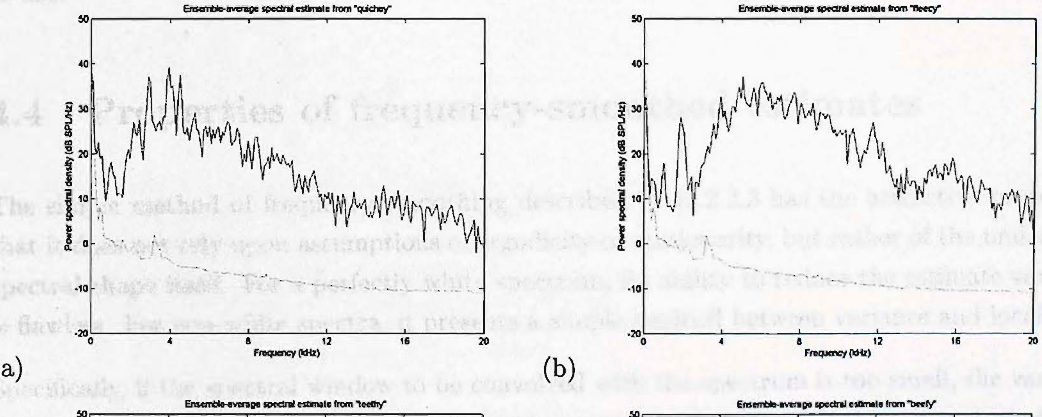


FIGURE 4.7: Ensemble-average spectral estimates using 6 windows from centre of separate fricative productions. (a) [ʃ] from centre of "quichey". (b) [s] from centre of "fleecy". (c) [θ] from centre of "teethy". (d) [f] from centre of "beefy" produced by M-01 .

A comparison of these ensemble-averaged spectra, with those formed by time-averaging in figure 4.5 certainly shows similarities in terms of overall energy distribution. However, a closer examination reveals that some of the most striking features of these spectra are not present in both estimates.

For example, while the time-averaged spectrum for /θ/ in Figure 4.5.c shows two distinctive peaks at approximately 2 kHz and 3 kHz, for the ensemble-averaged case in figure 4.7.c, the peak at 3 kHz has been obliterated. In the time-averaged estimate for /f/ in Figure 4.5.d, a double peak is the prominent feature around 2 kHz, but for the ensemble-averaged case in Figure 4.7.d, the double peak has also disappeared. The main peak magnitudes for /s/ and /ʃ/ in these two sets of plots also display contradictory information. For example, the 33dB peak at around 2 kHz in Figure 4.5.b has fallen to around 28dB in the ensemble average in Figure 4.7.b, and a new 'peak' at around 26dB has appeared in the 11.5-kHz region, where the time-averaged estimate showed a 19dB falling slope.

This is sufficient demonstration that observations based on estimates relying on assumptions that have not been proven, and may not hold well, may be misleading if interpreted without due consideration.

Finally, the practical value of ensemble averaging in situations where a particular production needs to be analysed is greatly limited. Only under considerably controlled circumstances can anything approaching a reliable ensemble be gathered. When studying the pathological speech of speakers who may have especially large variation in production, the method is almost impossible to use.

## 4.4 Properties of frequency-smoothed estimates

The simple method of frequency smoothing described in §2.2.2.3 has the attractive properties that it does not rely upon assumptions of ergodicity or stationarity, but rather of the underlying spectral shape itself. For a perfectly white spectrum, its ability to reduce the estimate variance is flawless. For non-white spectra, it presents a simple tradeoff between variance and local bias.

Specifically, if the spectral window to be convolved with the spectrum is too small, the variance will not be reduced sufficiently, and will remain unmanageably large. If the spectral window is too wide, local bias dominates: regions of the underlying spectrum with large first differential will be flattened, and peaks in the spectrum will lose their height, and gain width. Features lose their definition.

So how is the spectral window's shape and size determined? To answer this, knowledge of the underlying spectral shape is required, and again the paradox arises. This time, knowledge of the physical system may guide the choice.

Many of the considerations and aspects concerning the implementation of frequency-smoothing are given in the next section on multitaper analysis. For now, consider that the best bandwidth resolution (but highest variance of the estimate) that can be obtained using a sampling rate of 48 kHz, and 512-point data windows is $\Delta f = 93.75\,\text{Hz}$. The overlaid spectral estimates in figures 4.8 and 4.9 have been generated using rectangular spectral windows of $M = \{2, 4, 8, 16, 32\}$

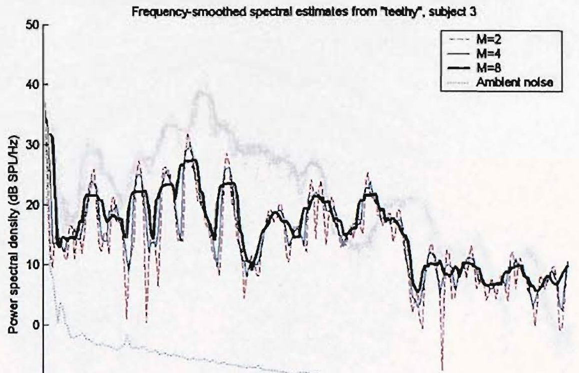


FIGURE 4.8: Frequency-smoothed spectral estimates using Daniell frequency windows from centre of [θ] in production of 'teethy' by subject M-03, for $M = \{2, 4, 8\}$ (top) and $M = \{8, 16, 32\}$ (bottom). Note that as a result of the frequency smoothing process, the very low frequencies cannot be shown. This effect also occurs at the high frequency end of analysis (up to 24 kHz), but the graph has been cut off at 20 kHz.

(bandwidths of 188, 375, 750, 1500 and 3000 Hz respectively). For small $M$, the variance dominates; for large $M$, the spectrum loses its shape. How the optimal value for $M$ will be decided is discussed in more detail shortly.

## 4.5 Benefits of multitaper analysis

As shown in §2.2.3, multitaper analysis is the optimal method of obtaining a reduced-variance estimate where a single short realisation of a stochastic process is required. It makes use of a larger amount of the original data, without introducing bias, while reducing the estimate variance by optimal use of an advanced form of frequency smoothing. The final representation is of tremendous value where accurate measurement of the spectrum within well-defined frequency-limits is required.

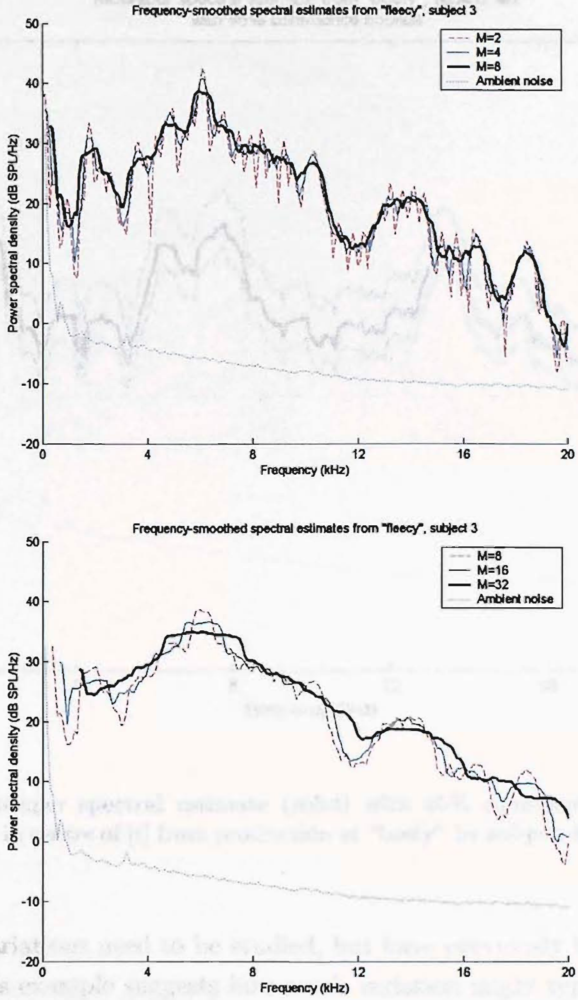


FIGURE 4.9: Frequency-smoothed spectral estimates using Daniell frequency windows from centre of [s] in production of 'fleecy' by subject M-03, for $M = \{2, 4, 8\}$ (top) and $M = \{8, 16, 32\}$ (bottom).

### 4.5.1 Reduced variance spectral estimates

Recall the estimate calculated using a modified periodogram in §4.1, shown in Figure 4.3. Using the same data, but analysing with the multitaper technique (10.6-ms, 512-point data windows, $NW = 4$) gives us the spectral estimate in Figure 4.10. Note that the 95% confidence interval is greatly reduced from that of Figure 4.3. Also notice the apparently sharp spectral peaks and troughs that have not been blunted by too-severe frequency smoothing.

A result of this reduced-variance spectral estimate is that actual changes in the underlying system (that were previously swamped by the variance of the estimate) can now be measured. As an example, it can be used to illustrate the change in peak position over time that we postulated was the cause of the double-peak in Figure 4.5.d (see §4.2).

Figure 4.11 shows multitaper spectral estimates of the six adjacent 10.6-ms windows that were used to form the ensemble averaged estimate in Figure 4.5. While appearing rather cluttered, the plot clearly shows the gradual change over time in the frequency and amplitude of the peak

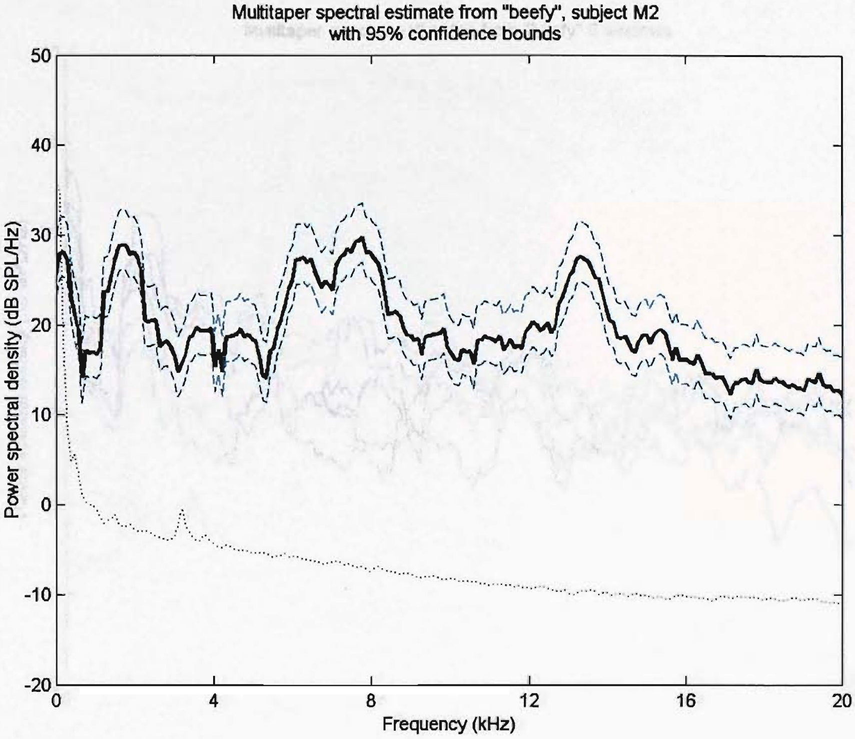


FIGURE 4.10: Multitaper spectral estimate (solid) with 95% confidence bounds (dashed). Taken from 10.6 ms in centre of [f] from production of "beefy" by subject M-02 (cf. Figure 4.3).

around 2 kHz. Such variations need to be studied, but have previously been ignored as a result of time-averaging. This example suggests how much variation might typically be expected over a 64-ms interval of a fricative. It also highlights how much of this variation has been previously swamped by the high variance in poor spectral estimates, and not captured by broad measures such as spectral moments.

### 4.5.2 Comparison to frequency smoothing

In §2.2.3 it was shown that the power frequency response for the combined multitaper transform is a wide, flat-topped main-lobe with small trailing side-lobes. It was demonstrated that a similar response is obtained by convolving the power frequency response of say, a Hanning window with a rectangular frequency window: essentially the process undertaken during frequency smoothing.

A comparison can now be made between multitaper estimation and frequency smoothing. Since the multitaper estimate makes use of a larger proportion of data within each window, and since the prolate spheroidal sequences are optimal at maximising energy while minimising bandwidth, the multitaper estimate should always be considered the best estimate where frequency tradeoff methods are being compared.

Spectra from central /ʃ/ and /s/ spectra are shown in figures 4.12 and 4.13, for smoothed periodogram spectra for $M = \{2, 4, 8\}$, and in the larger figures 4.14 and 4.15 for $M = 6$, superimposed on multitaper spectra with $NW$ bandwidth product of 4. An approximation to

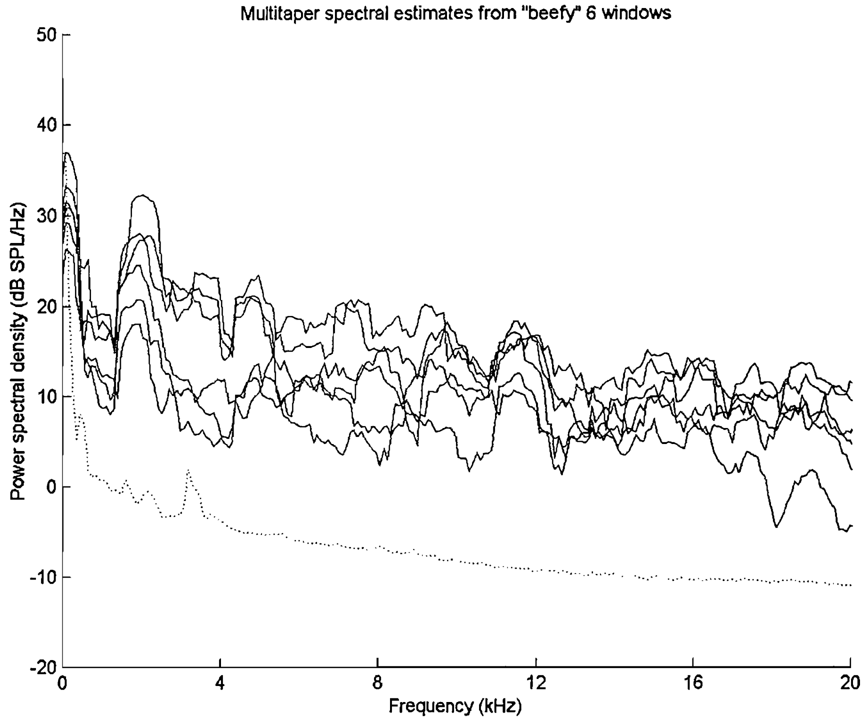


FIGURE 4.11: Superimposed multitaper spectral estimates over 6 adjacent 10.6-ms time windows, from /f/ in 'beefy', subject M-01. These same data windows were used to form the time-averaged estimate in Figure 4.5.d. Note the change over time of the main peak at around 2 kHz.

the main lobe of the response of combined multitaper analysis with $NW = 4$ bandwidth product is to set $M = 4$ in the convolution of the modified periodogram spectrum with a Daniell window. The resolution of the spectrum without smoothing is $f_s/512 = 94$ Hz. With $M = 4$ smoothing, this resolution is reduced to $Mf_s/512 = 375$ Hz. While the responses look similar in terms of their main-lobe widths, the multitaper analysis reduces the variance by a factor of nearly 7, while frequency smoothing with $M = 4$ will only reduce the variance by a factor of 4 in regions of the underlying spectrum that are flat. This explains why many regions of the frequency-smoothed estimates in figures 4.12.b and 4.13.b still appear more 'spiky' than the multitaper estimates, despite having similar resolutions.

Setting $M = 6$ (as demonstrated in figures 4.14 and 4.15) in the frequency-smoothed spectrum apparently provides a closer estimate to that of the multitaper spectrogram (and, we assert, to the true spectrum). However, a few points are worth mentioning. Firstly, the increase in $M$ means an increase in the main lobe width, and this results in a further reduction in the spectral resolution to $Mf_s/512 = 563$ Hz. This means that two frequency points that lie closer together than this are heavily correlated, and cannot be used reliably.

Secondly, the local bias is increased. The $M = 6$ smoothed spectrum generates a better approximation to the multitaper spectrum in the main peak of Figure 4.14 by reducing large variances in this region. However, a problem arises in using a fixed value of $M$, which may give a good spectral estimate for one segment of data, but may not be the best choice for another. This can be demonstrated by considering the suppression of the small energy peak around 3 kHz in figures

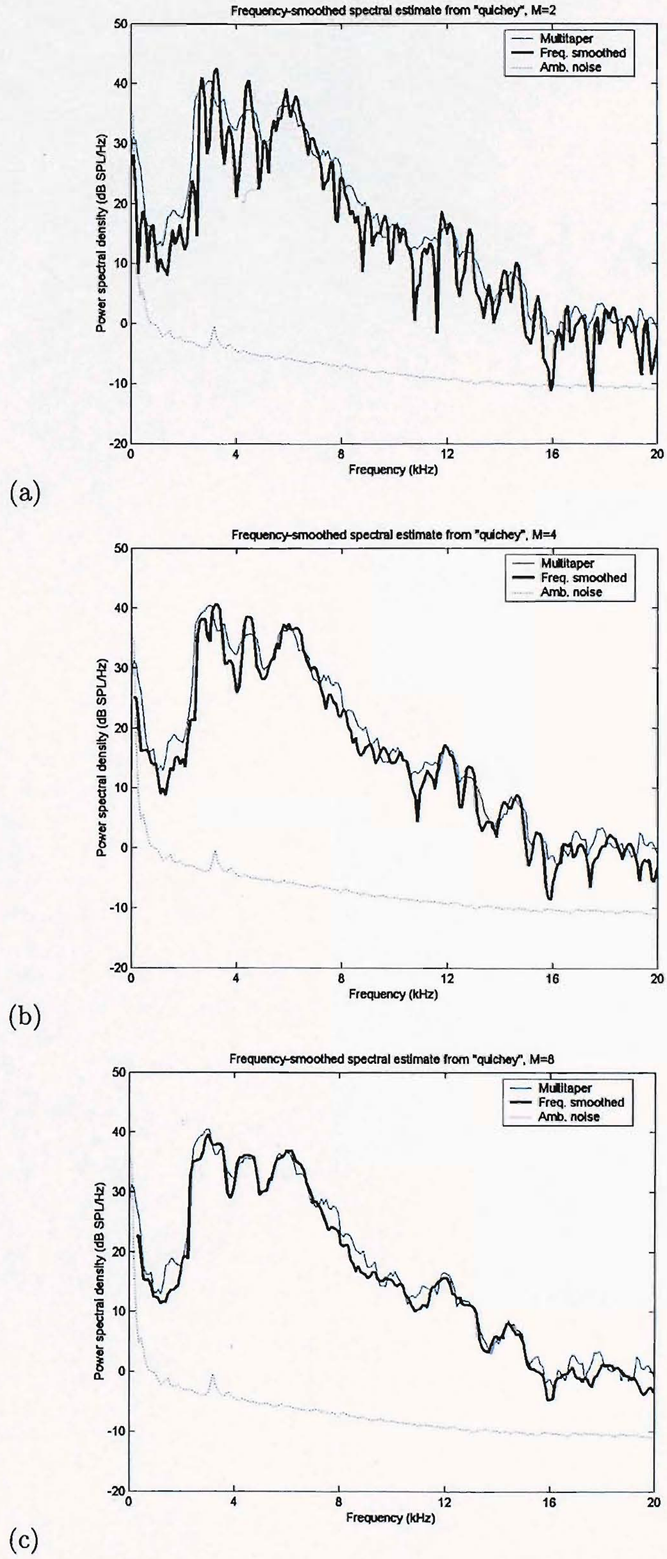


FIGURE 4.12: Multitaper spectral estimates with frequency-smoothed estimates superimposed for $M = \{2, 4, 8\}$ for (a), (b) and (c) respectively. 10.6-ms section taken from centre of [ʃ] from production of "quichey" by subject M-02.

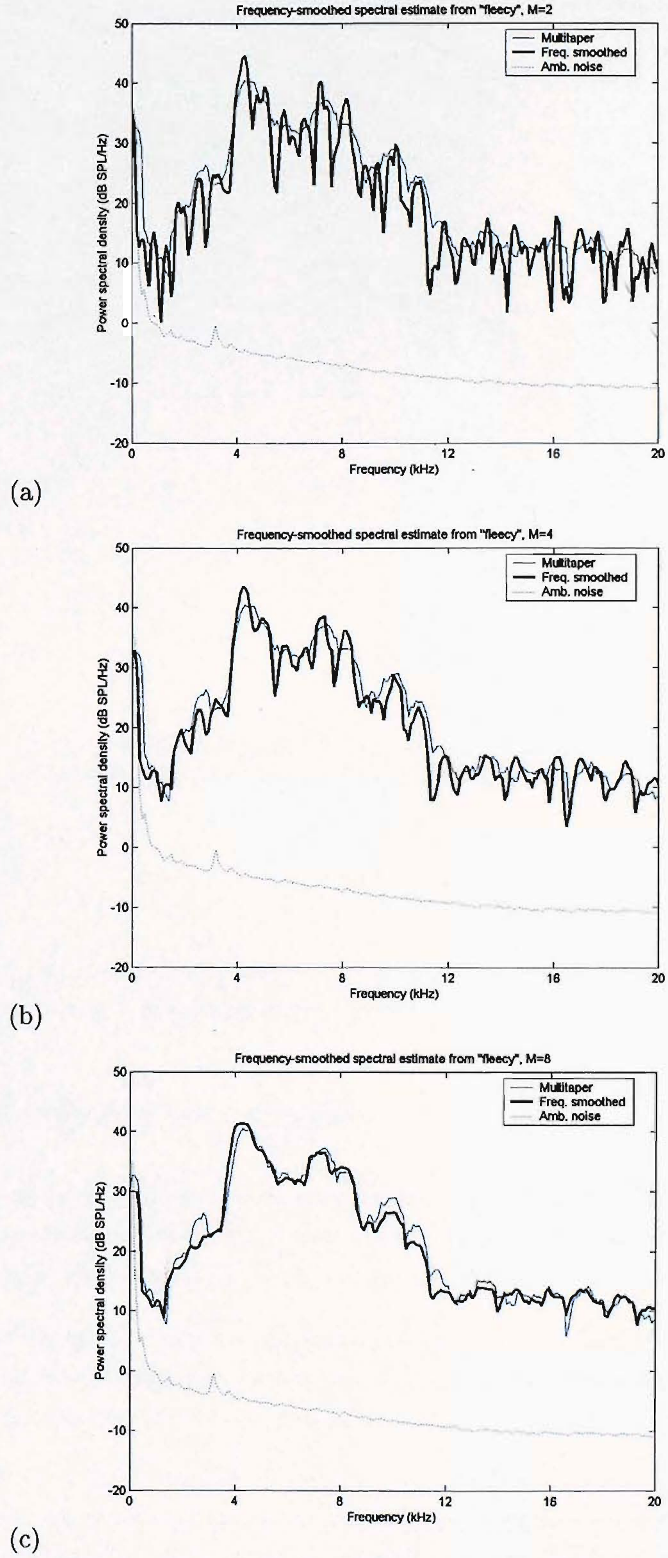


FIGURE 4.13: Multitaper spectral estimates with frequency-smoothed estimates superimposed for $M = \{2, 4, 8\}$ for (a), (b) and (c) respectively. 10.6-ms section taken from centre of [s] from production of "fleecy" by subject M-02.

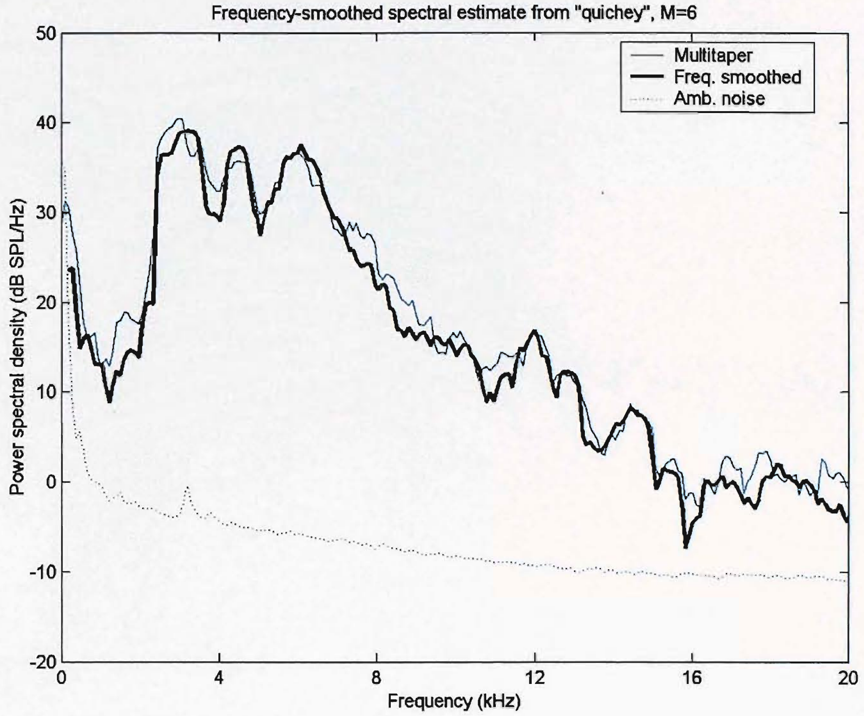


FIGURE 4.14: Multitaper spectral estimates with frequency-smoothed estimates superimposed for $M = 6$. 10.6-ms section taken from centre of [ʃ] from production of "quichey" by subject M-02.

4.13.b and 4.15.

Further increasing the size of the rectangular frequency window to $M = 10$ results in an estimate that is starting to seriously bias spectral features.

### 4.5.3 The multitaper spectrogram

Multitaper spectral estimates have been shown to be very reliable by minimising estimate variance, while maintaining low local bias (compared to other frequency-smoothing methods), and without relying upon assumptions of ergodicity or stationarity.

Unlike ensemble-averaged, or time-averaged spectral estimates, multitaper estimates can be used in the construction of spectrograms. Such a spectrogram has been constructed for the same dataset as that used to generate Figure 4.4, and is shown in Figure 4.16.

At once the differences between these two interpretations of the same data can be assessed. The large variance of the original estimate gives the fricative portion a coarse appearance, while the multitaper estimate is far more 'predictable'. Indeed, the new representation is the one desired, since it reflects the fact that the underlying production mechanism is not varying rapidly, but is smooth.

Consider now the representation by this multitaper spectrogram of the vowel structure on either side of the central fricative. A result of the spectral smoothing is that the very fine formant

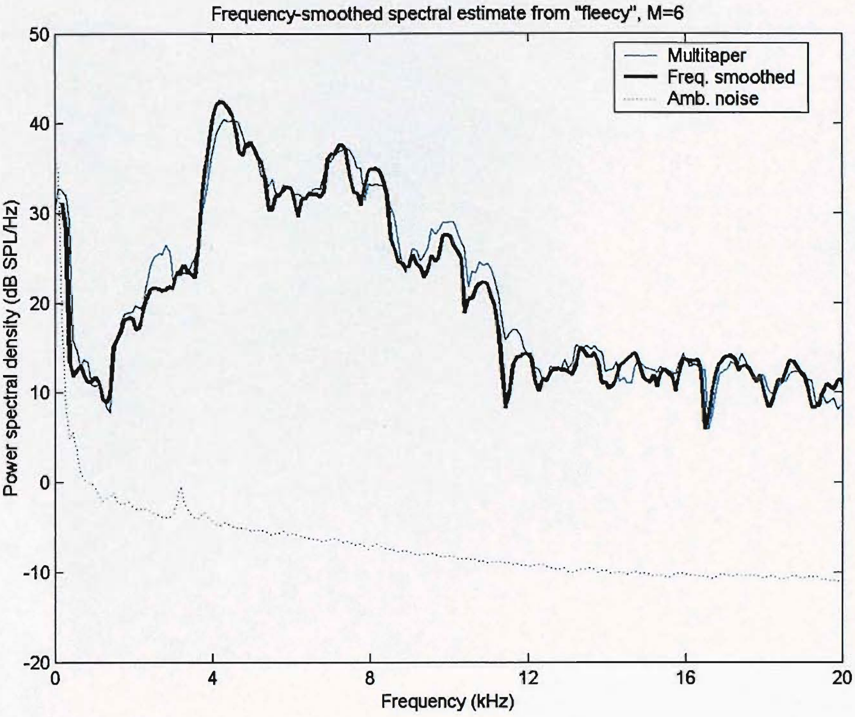


FIGURE 4.15: Multitaper spectral estimates with frequency-smoothed estimates superimposed for $M = 6$. 10.6-ms section taken from centre of [s] from production of "fleecy" by subject M-02.

structure of the vowels on either side of the fricative has lost its definition. This is strong demonstration that the multitaper technique, while well-suited to fricative analysis, is the incorrect choice for analysing vowels. The computations within the multitaper procedure are completely unnecessary for the study of deterministic and pseudo-deterministic signals such as vowels: no extra information is acquired, and spectral resolution is lost.

Spectrograms constructed using multitaper spectral estimates, are presented for all voiceless fricative productions from the normal-hearing speakers, in Appendix B. These spectrograms include information up to 20 kHz, and demonstrate the large variations that exist across tokens.

## 4.6 Summary

The estimates commonly used in the literature are subject to large variances, which can obscure spectral features, and present difficulties for peak-tracking.

The methods for attempting to reduce estimate variance have been shown to sometimes be problematic, and cause limitations in the aspects of data that can be analysed: time-averaging prevents changes over time from being measured, while ensemble-averaging prevents changes over production from being measured.

Multitaper analysis allows a reduced-variance spectral estimate for a single window of fricative data to be generated, by making use of more information than a single periodogram, and repre-

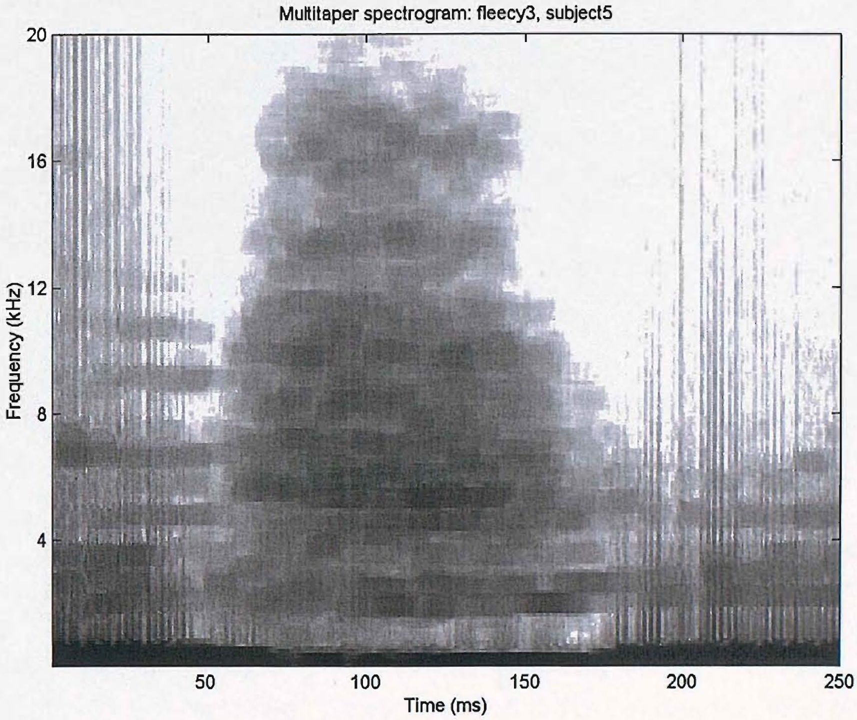


FIGURE 4.16: Multitaper spectrogram of [isi] production from 'fleecy' by subject M-05. $N =$ 512, overlap= 480 (cf. standard spectrogram in Figure 4.4).

senting this at an optimal resolution by maximising the energy while minimising the bandwidth. While frequency smoothing often produces comparatively similar results, it does not do so in a reliable manner. That is, the best results from frequency-smoothing come after $M$ has been adjusted with reference to the multitaper spectral estimate. Even so, the multitaper spectrograms excel where short windows of time-series data are to be examined, since a greater proportion of data are incorporated into the estimate.

Since multitaper analysis does not rely upon assumptions of stationarity, it is now possible to study with greater accuracy changes in the spectrum over time. Because no ensemble-averaging takes place, differences between productions can be studied. These features are of great value, and are investigated in detail in Chapter 6.

# Chapter 5

# Analysis: spectral moments

Spectral moments have become one of the most popular methods for characterising fricative productions. They have been used to measure differences in productions of the same fricative across and within subjects. They have been found to be beneficial in comparisons of normal and disordered speech, and also for measuring changes in disordered production that may occur over long periods of time after the hearing system has been significantly changed.

However, in some instances they have been found to produce inconclusive results. Their ability to discriminate between the voiceless sibilants has generally been found to be good; however, they have so far not been able to distinguish the non-sibilant voiceless fricatives.

The performance of spectral moments is examined. Several parameters that have up until now been chosen rather arbitrarily, are considered. Adjustments that may result in small improvements in the performance of spectral moments are presented in §5.1.

It has often been considered that significant cues may lie in the spectral changes that occur over time. Tracking the changes in spectral moments of fricative productions over time has been attempted in a few places in the literature. Typical variations over time of spectral moments are examined in §5.2.

Finally, a discussion of the findings is presented.

## 5.1 Adjustments to spectral moments

A number of parameters inherent in the spectral moment methodology require careful consideration. These parameters can generally be chosen arbitrarily, and still produce an apparently satisfactory set of results. However, by more careful consideration of these parameters, it may be possible to improve the performance of spectral moments somewhat.

Two main criteria are open to improvement. Spectral moments are employed to capture the *shape* of a spectral distribution. Each moment describes a particular characteristic of a distribution, and if the methodology has been properly optimised, each moment should be sensitive to a single aspect of the spectral shape; moreover, correlation across the moments should be minimised. This

| | Multitaper | Periodogram |
|---|---|---|
| Centroid | $7.8 \times 10^3$Hz | $7.5 \times 10^3$Hz |
| Variance | $24.9 \times 10^6$Hz$^2$ | $23.5 \times 10^6$Hz$^2$ |
| Skewness | $73.2 \times 10^9$Hz$^3$ | $75.5 \times 10^9$Hz$^3$ |
| Kurtosis | $-318.4 \times 10^{12}$Hz$^4$ | $-193.2 \times 10^{12}$Hz$^4$ |

TABLE 5.1: Moments generated from multitaper and periodogram spectra of example [ʃ] token.

is achieved by ensuring that the distribution to be characterised by spectral moments is almost Gaussian. If the distribution does not resemble a Gaussian distribution, then the moments will become less sensitive to spectral shape, and more highly correlated to each other. The closer the distributions are to Gaussians, the better the spectral moments will perform, and hence the better their distinguishing capabilities.

All spectral moments are calculated from a single 10.6-ms mid-fricative data window, unless otherwise stated. The term 'fricative token' is used to mean a single, mid-fricative 10.6-ms data window.

### 5.1.1 Spectral moments of multitaper spectra

Spectral moments have until now used modified periodogram spectral estimates over some portion of the fricative under examination. It is known that the spectral estimates resulting from a simple modified periodogram are subject to large variance error. In some instances, time or ensemble averaging techniques have been used in attempts to reduce these errors. However, it remains unclear as to whether fricative production satisfies the conditions necessary for the averaging techniques to produce reliable, informative results.

Multitaper analysis produces consistent spectral estimates, with reduced variance error. These spectral estimates can therefore be used to calculate the spectral moments

Keeping all other factors constant, the spectral moments calculated from each type of spectral estimate are compared. Frequency scales are up to 20 kHz unless otherwise stated. As a first example, consider the centre of an [ʃ] token in an /iʃi/ context. Periodogram, and multitaper spectral estimates for such a token, from male subject M-02 are shown in Figure 5.1. The moments calculated from the multitaper spectrum and from the periodogram spectrum are shown in table 5.1.

We first note that the minimum values in the periodogram spectral estimate lie within only a few decibels of −20dB SPL/Hz, meaning that, at least for this token, the zero reference should not be set higher than about −20dB SPL/Hz. The multitaper spectrum however, shows that the zero reference could be set substantially higher, around −5dB SPL/Hz, without any clipping of the spectrum occurring. Setting the zero reference is discussed more fully in §5.1.3.

The Gram-Charlier distributions also indicate which features of the spectrum the spectral moments are most sensitive to: the broad peak position, and the fall-off of the tails have influenced the moments strongly, but the finer spectral peaks shown most clearly by the multitaper spectrum at around 4 kHz and 7 kHz have not had much 'influence'.

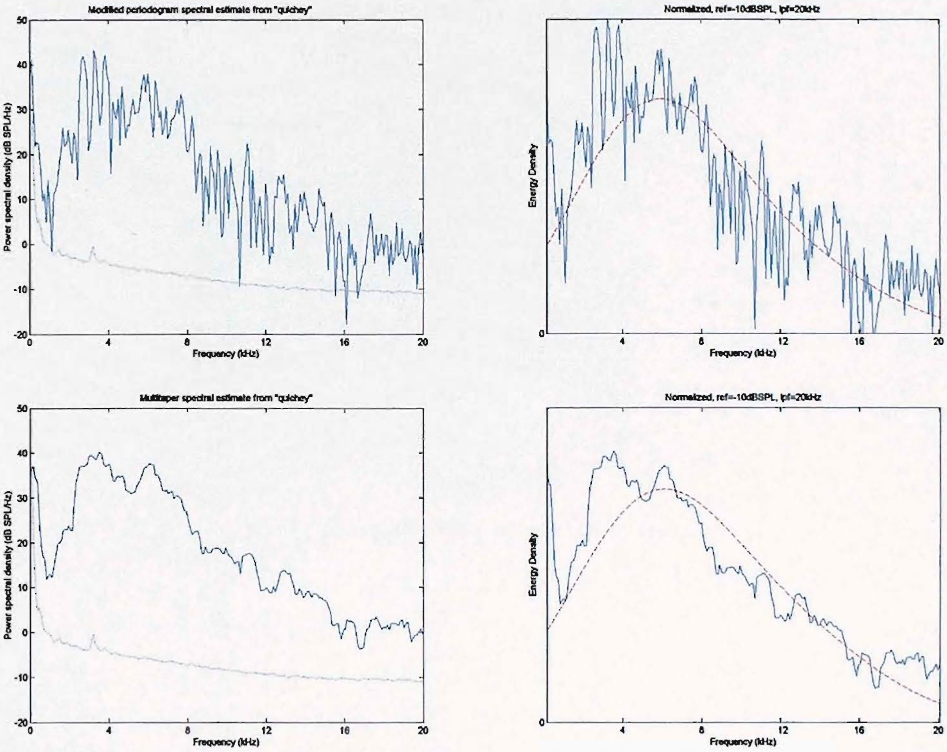


FIGURE 5.1: 512-point modified periodogram (top) and multitaper (bottom) spectral estimates from centre of [ʃ] in /iʃi/, left, and right, normalised and with Gram-Charlier distribution (smooth curve) from corresponding spectral moments calculated with −10dB SPL/Hz zero-reference, and 216 Hz–20 kHz frequency range. Dotted lines on left show ambient room noise.

Comparing the moments (given in table 5.1) from the multitaper spectrum to those calculated using the periodogram spectrum, it can be seen that the low-order moments are fairly similar, and this is reflected in the similar Gram-Charlier reconstructed distributions. This is to be expected, since they are basically averaging operators over the whole spectrum, so that individual estimation errors are 'smoothed' out. This makes the low-order spectral moments good descriptors of spectral distributions with large variance across their frequency range, and so are quite well-suited to the task of describing spectra calculated from crude estimates.

The skewness operator is simply a measure of which side of the spectrum has greater mean energy, and so can also be expected to give similar results. However, for the fourth-moment, (and for higher-order moments), small variations in the tails of the distribution will become exaggerated. Since the errors in the tails of the periodogram spectrum are much greater than for the multitaper spectrum, larger variations of values for these high-order moments would be expected.

We wish to examine how significant the variations due to spectral error variance in calculated spectral moments are, compared to typical variations across productions. Figure 5.2 compares the 3rd and 4th spectral moments calculated from periodogram, and multitaper spectral estimates. Each set of moments is calculated from a single 10.6-ms mid-fricative data segment. The voiceless sibilants, in all vowel contexts, as produced by all male subjects, are presented. Blue plus-signs indicate /s/, red circles /ʃ/. Figure 5.2 suggests that the improved multitaper spectral

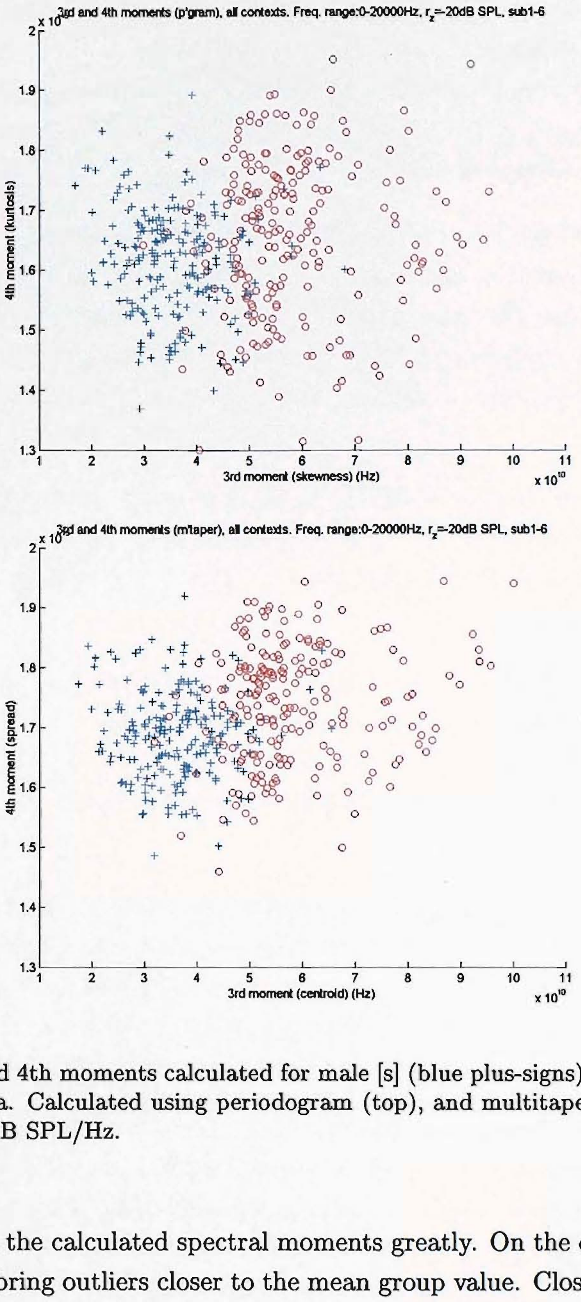


FIGURE 5.2: 3rd and 4th moments calculated for male [s] (blue plus-signs) and [ʃ] (red circles) mid-fricative spectra. Calculated using periodogram (top), and multitaper (bottom) spectral estimates, $r_z$=−20dB SPL/Hz.

estimate does not alter the calculated spectral moments greatly. On the occasions when it does, the effect is usually to bring outliers closer to the mean group value. Closer analysis reveals that the most extreme outliers are caused by spectra which exist below the set zero reference at some frequencies; the reduced variance of the multitaper estimates means this is less common, and so the moments become slightly more stable.

Using multitaper spectral estimates to calculate spectral moments means that the zero reference can potentially be set higher, thus increasing the sensitivity of the spectral moments, while reducing moment variations due to error in the spectral estimate. In §5.1.3 the effects of varying the zero reference prior to calculation of spectral moments are explored.

### 5.1.2 Appropriate frequency-range selection

Fricative studies have traditionally focussed on frequencies below about 10 kHz. There have been a number of motivational reasons for limiting the frequency analysis methods to around half the human perceptual range. Certainly it is well established that human perception of fricatives remains high when the signal above, say, 8 kHz is filtered out; hence, limiting the search for primary discriminatory cues to this frequency scale seems justifiable. Nevertheless, studies have shown that perception cues exist well above 8 kHz (Lippmann 1996), suggesting that the production mechanism is generating cues at higher frequencies.

In the fricative perception analysis of cochlear implant users, who generally can only make use of information up to a maximum of around 5 kHz, comparisons to perceptual capabilities of normal-hearing subjects presented with similar frequency-limited signals is clearly warranted. However, when studying the production of fricatives by such subjects, there is no reason not to suspect changes in the higher frequency regions of the signals, where potential cues may lie, and which may indicate subtle changes in production.

Analysis of fricative production should not begin by discarding information at frequencies that are not often required for good perception classification. Indeed, some interesting production characteristics can be quickly observed by considering just a few multitaper spectra of typical fricative productions.

An example mid-vowel spectrum of [i] from a production of the word "b<u>ee</u>fy" is presented in Figure 5.3 (top). Note that energy at all frequencies, including those above 11 kHz, is at least 10dB above ambient room noise, and has a mean of approximately 5dB SPL/Hz in the 10–20-kHz frequency range, indicating that indeed, the production mechanism is producing energy in the higher frequency ranges.

The middle plot in Figure 5.3 shows the mid-fricative [f] spectrum of the same production. It is not difficult to notice that the large proportion of energy in this fricative occurs above 6 kHz, at a mean level of approximately 18dB SPL/Hz in the 6–20-kHz frequency range. Significantly, these energy levels are also greater than for the vowel spectrum, suggesting that this is possibly an important aspect of the fricative production.

Finally, the bottom plot in Figure 5.3 shows a spectral slice on the [fi] fricative-vowel boundary, where the energy has dropped in the high-frequency range to lower levels than any other point during vowel or fricative production, almost dropping to ambient room noise levels above 11 kHz. This short temporal occurrence may also reflect another cue characteristic of the fricative, that could only be captured if the frequency range used is great enough.

Using data from all male subjects, the effects of varying frequency range on spectral moment sensitivity is demonstrated in Figure 5.4. The first and second spectral moments for all voiceless fricatives in all vowel contexts as produced by all male subjects are presented.

It is inevitable that we will occasionally need to quantify the 'separation' of such multivariate fricative clusters. This will clarify any degree of improvement in fricative cluster separability when using different descriptive parameters. Tests of certainty that distributions have different means (such as p-tests) are insufficient for the purposes of describing the degree by which two clusters of tokens are separated. A more suitable method for quantifying the separation of two clusters is Fisher's linear discriminant, given

$$J = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \cdot \boldsymbol{d}}{\boldsymbol{d}^T \cdot (V_1 + V_2) \cdot \boldsymbol{d}} \tag{5.1}$$

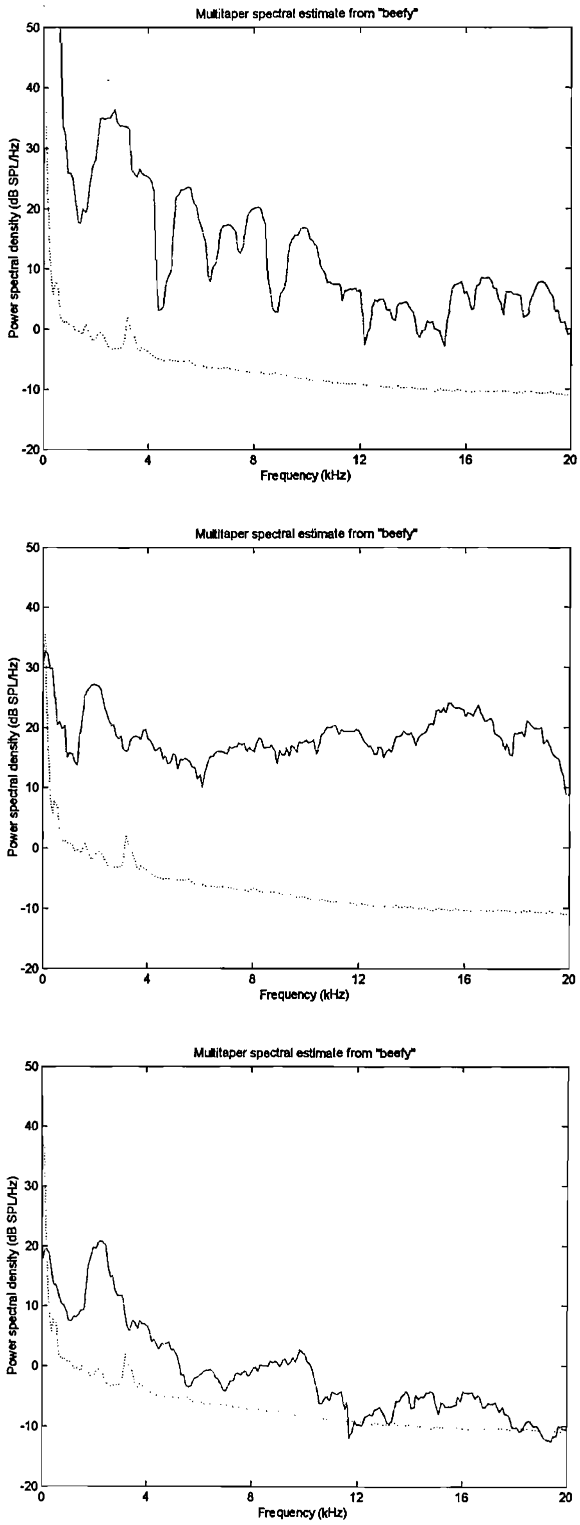


FIGURE 5.3: Mid-vowel [i] spectrum (top), mid-fricative [f] spectrum (middle), and [fi] fricative-vowel boundary spectrum (bottom), 10.6-ms data windows from a single production of "beefy", subject M-01. Dotted line is ambient room noise.

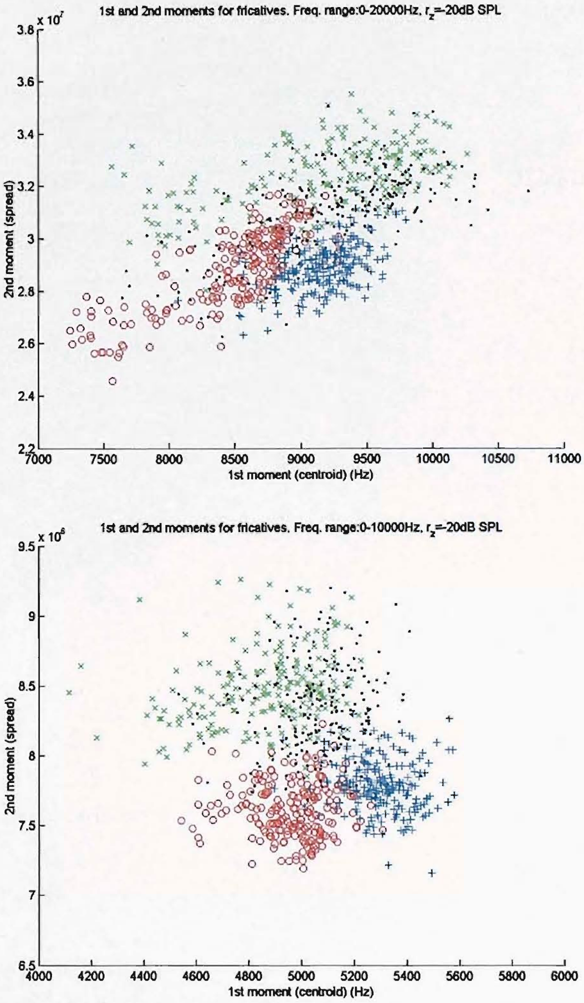


FIGURE 5.4: Spectral moments of all voiceless fricatives in all contexts across all male subjects; calculated from single, mid-fricative 512-point multitaper estimates using 216 Hz–20 kHz range (top), and 216 Hz–10 kHz range (bottom), $r_z$=−20dB SPL/Hz. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

where $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the means of each fricative cluster, $\boldsymbol{V_1}$ and $\boldsymbol{V_2}$ are the covariance matrices of each fricative cluster, and $\boldsymbol{d}$ is the direction vector between the two clusters. The value $J$ represents the distance between the two clusters, in terms of the variances of each cluster. Generally a value of $J \gtrsim 4$ indicates that the clusters are well separated, while values of $J \lesssim 2$ indicates that the clusters are well within two standard deviations of each other, and hence are likely to be overlapping.

Returning to Figure 5.4, notice that when the 20-kHz frequency range is used, the first and second spectral moments appear to be weakly correlated ($C = 0.58$, where $C$ is the coefficient of correlation). The effect of reducing the frequency range from 20 kHz to 10 kHz is more distinct however, and of importance is the reduction in correlation of first and second spectral moments (to $C = -0.22$). The separation of the sibilant clusters has also increased from $J = 3.19$ to $J = 3.94$. This slight reduction in the apparent correlation between the first two moments, has resulted from a rather arbitrary change in frequency range selection. Knowledge of expected

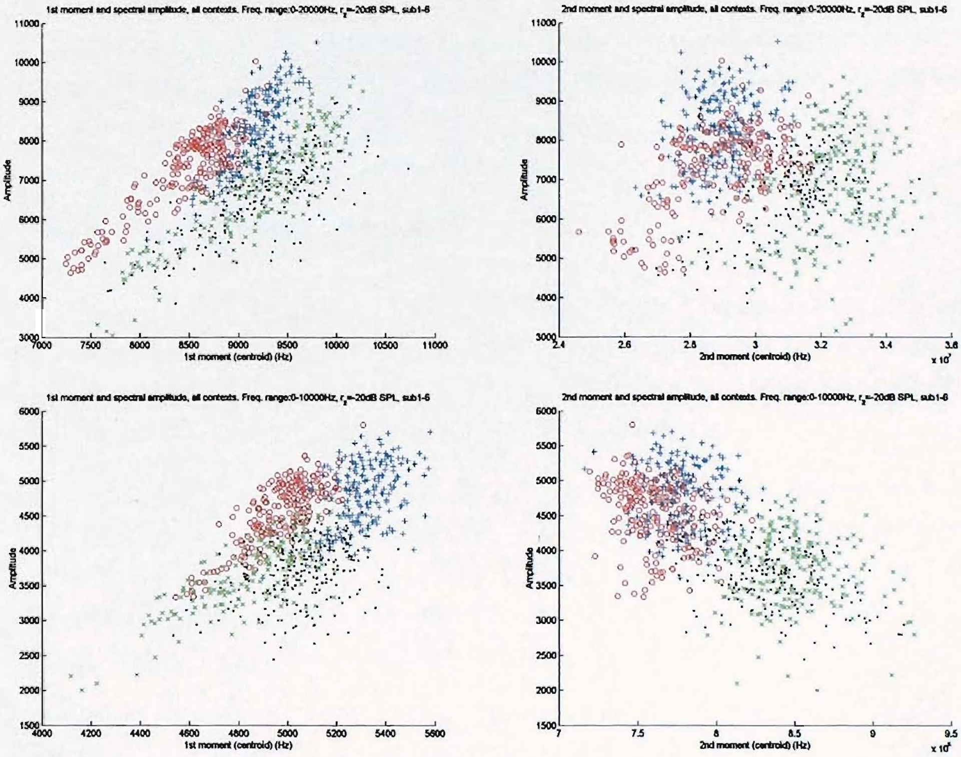


FIGURE 5.5: Plots of 1st (left) and 2nd (right) spectral moments against total spectrum amplitude. All voiceless fricatives, all contexts, all male subjects. Calculated from multitaper spectra over 0–20 kHz (top) and 0–10 kHz (bottom). $r_z$=-20dB SPL/Hz. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

production spectra has not been used in determining an optimal frequency scale. In fact, the reduced correlation of the lower plot partially results from a reduced sensitivity of the spectral moments. The reduction in frequency scale has not been accompanied by a suitable change in zero reference, and so the 'tails' of the spectral distributions are generally higher than for the 0–20-kHz frequency range. Spectral moments will be insensitive to changes in distributions with high tails, and so changes in frequency range selection must be accompanied by careful selection of the zero reference.

Another correlation can be seen by plotting the first and second spectral moments against total spectrum magnitude before normalisation. Figure 5.5 demonstrates these correlations for 0–20-kHz, and a 0–10-kHz frequency ranges. Notice the moderate correlation of the first moment to total spectral amplitude ($C$ = 0.56 for the 0–20-kHz range, and $C$ = 0.61 for the 0–10-kHz range). Of course, this is to be expected, due to the fixed nature of power normalisation. As the overall intensity of the fricative increases, the spectral shape moves up the $y$-axis, further from the zero reference. After normalisation, the higher distributions will thus inevitably appear 'flatter'. Since /s/ spectra generally have significantly more energy at lower frequencies, the normalisation process will effectively redistribute the energy of louder productions to the right, and hence the first moments of /s/ are particularly correlated. The second moments are loosely negatively correlated when the 0–10-kHz frequency range is used ($C = -0.64$), although very little overall correlation is found in the second moments when using the 0–20-kHz range ($C = -0.05$).

It may be that correlations exist between the spectral shape of a fricative, and its total spectral amplitude. However, it is almost impossible to determine whether such a correlation exists using spectral moments, due to the *expected* correlations imbedded in their calculation, when considering distributions with high tails.

### 5.1.3 Selection of zero-reference

One of the first steps needed to calculate the moments of a spectral distribution, is to determine a suitable zero reference to use when normalising the spectrum so that it can be treated as a distribution with total area equal to unity (as discussed in §2.3). It is worthy of note again that the literature seems to make no mention of setting the zero reference.

A number of practical considerations force us to choose a somewhat arbitrary zero reference value. For instance, it may be chosen as some fixed number of decibels below the spectral peak, or even at the ambient noise floor. While these arbitrary decisions may eliminate certain problematic variables, they also tend to create several new ones. It is important to consider the effect upon the spectral moment methodology itself, when choosing the zero reference.

To ensure that the spectral moments are sensitive to the shape of the power spectrum, the distribution must approximately resemble that of a Gaussian curve. That is, the tails of the distribution should be close to zero, and it should comprise of a single, broad peak. If this is not the case, the resulting spectral moments will be insensitive to the changes in spectral shape that we are attempting to capture, which in turn will lead to difficulties in the interpretation of results.

Ideally then, spectra would be considered from a zero reference close to the low and high-frequency tails of spectrum. However, another of the most important requirements is that no point in the distribution lies below the zero-reference, since this would present problems when the normalisation step is performed. The distribution must be positive everywhere, or the resulting calculated moments will become nonsensical. Given the typical range of variations in fricative production, it is straightforward to demonstrate that satisfying both of these conditions simultaneously, using a fixed zero reference for a large set of tokens, is almost impossible.

In fact, fixing the zero reference around −10dB SPL/Hz was found to be approximately the highest value of zero reference that would ensure no multitaper voiceless fricative spectra ventured below $r_z$, using a frequency range of 0–10 kHz, across all our male and female subjects, and across all voiceless fricative tokens (for modified periodogram spectra, $r_z$ had to be set to −20dB SPL/Hz in order to achieve a similar amount of stability). If set higher than this, the zero reference starts to seriously clip the spectral data with increasing regularity, causing the spectral moments to become increasingly unstable, resulting in more and more outliers.

Unfortunately, many fricative tokens have spectra well above this reference. On normalisation of these spectra, the tails of the distribution do not approach zero, and hence, the spectral moments become less sensitive to changes in these regions. An undesirable side-effect is that the odd-order moments become highly correlated to each other, as do the even-order moments.

To take an example, consider the multitaper spectrum shown in the bottom-left of Figure 5.1. Consider the frequency range 800 Hz to 12 kHz. This shape can be described well by a Gaussian curve, so long as the zero-reference is sufficiently close to the 'tails', as demonstrated in Figure 5.6. The moments calculated from the curve (see table 5.2) normalised using $r_z$ = 10dB SPL/Hz are fairly well suited using this zero-reference: the tails of the resulting distribution are close to 'zero'. The curve normalised using $r_z = -10$dB SPL/Hz of course appears much 'flatter', and the moments and Gram-Charlier distribution reflect this. The moments will have reduced sensitivity, since variations in the distribution will seem less significant than the equivalent variation using $r_z$=10dB SPL/Hz.

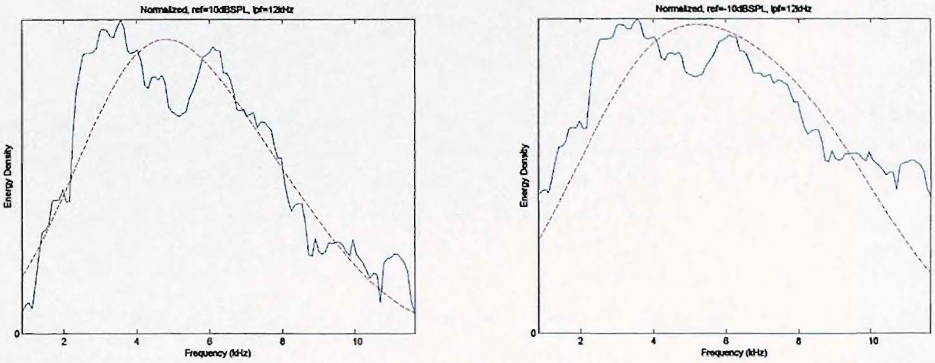


FIGURE 5.6: Normalised spectrum from centre of [ʃ] from "quichey" with Gram-Charlier curves calculated from corresponding moments, $r_z$ = 10dB SPL (left), $r_z = -10$dB SPL (right).

| | $r_z$ = 10dB SPL | $r_z = -10$dB SPL |
|---|---|---|
| Centroid | $5.5 \times 10^3$Hz | $5.9 \times 10^3$Hz |
| Variance | $6.1 \times 10^6$Hz$^2$ | $8.3 \times 10^6$Hz$^2$ |
| Skewness | $6.4 \times 10^9$Hz$^3$ | $5.1 \times 10^9$Hz$^3$ |
| Kurtosis | $-21.1 \times 10^{12}$Hz$^4$ | $-67.6 \times 10^{12}$Hz$^4$ |

TABLE 5.2: Spectral moments generated from multitaper spectra of example [ʃ] tokens, with different zero-references (see Figure 5.6).

However, pushing the zero reference above $-10$dB SPL starts to produce a few spurious results: spectra which exist below the zero reference result in extreme outlying sets of moments. Setting $r_z$ to 0dB SPL/Hz, a greater proportion of the spectra do not cope well with the normalisation procedure. With increasing $r_z$, the number of spurious results grow faster than the slight benefits from the better fit occasionally produced, and so little is gained by pushing the zero reference too high.

Figure 5.7 demonstrates the high degree of *structural* correlation across the odd and even order moments, using a 0–20-kHz frequency range, and $r_z = -20$dB for all voiceless fricatives produced in all vowel contexts by all female subjects. The coefficient of correlation for the first and third moments are $C = -0.97$, and for the second and fourth moments $C = 0.96$. This correlation is a result of the methodological approximations that have to be conceded when generating spectral moments.

It may seem that, since a fixed zero reference produces moments that are often either insensitive, or unstable, an appropriate step would be to set it as some function of the spectral peak amplitude, or perhaps the mean spectral amplitude. However, it is clear that variables such as the

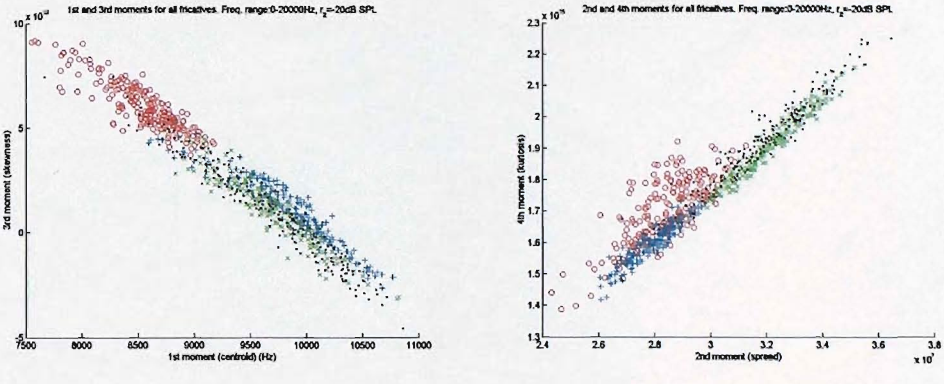


FIGURE 5.7: 1st against 3rd (left), and 2nd against 4th (right) spectral moments for all female subjects, all contexts, all voiceless fricatives. From multitaper spectral estimates over frequency range 0–20-kHz, $r_z$=-20dB SPL/Hz. Note the high degree of correlation. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

spectral peak, and the mean spectral amplitude are highly nonlinear functions of the amplitude at the tail frequencies of the distribution; such a function could not be made to produce stable results.

Clearly, the amplitude of the spectrum (or zeroth moment) should be recorded, in order to evaluate whether correlations between total spectral amplitude and spectral moments can be used to improve the separation of the spectral moments. Figure 5.8 attempts to plot 0th, 1st and 2nd moments in a three-dimensional plot, to see if the incorporation of total amplitude aids separation. If the plot in Figure 5.8 is rotated, it can be seen that even in this three-dimensional space, the non-sibilants completely overlap. Including the total spectral amplitude data does not improve the separation of the spectral moments.

The main problem is that the spectral shape of fricatives is highly variable, and often does not assume the form of anything resembling a Gaussian curve. A fine example of this can be seen in the mid-/f/ spectrum in Figure 5.3. This very flat spectral shape is common amongst the non-sibilants, and this is clearly one of the root causes of the inability of spectral moments to differentiate them.

### 5.1.4 Summary

The reduced variance error of multitaper spectral estimation allows the zero reference to be raised, without risking spectral clipping. The closer to the tails of the distribution the zero reference can be raised, the more sensitive the spectral moments will be to changes in the spectral shape. Nevertheless, it has been found that a zero reference of −10dB is most satisfactory for minimising the number of spurious spectral moment values.

A high degree of correlation exists amongst even moments, and amongst odd moments (demonstrated in Figure 5.7). These correlations cannot be reduced by raising the zero reference above around -10dB SPL/Hz (due to the growing number of spurious sets of moments resulting from the sheer variability across productions), it becomes apparent that the first two moments are likely to yield as much information as any other combination of moments.

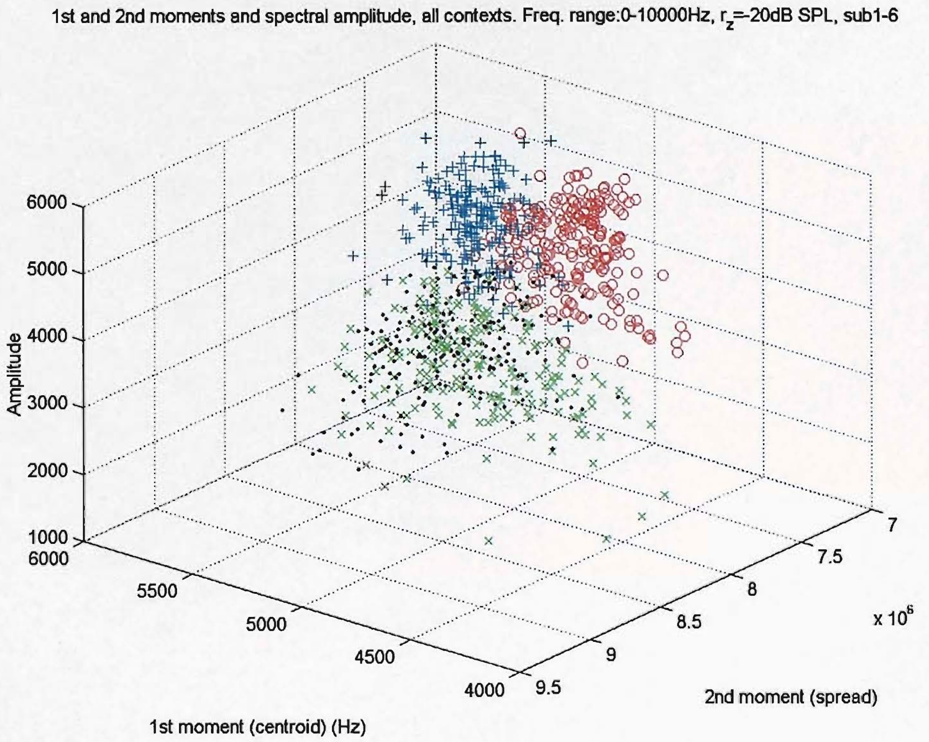


FIGURE 5.8: 3D plot of amplitude, 1st, and 2nd moments, all voiceless fricatives, all male subjects, all vowel contexts. Frequency range: 0–20-kHz, $r_z$=-20dB SPL/Hz. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

While variations in frequency range potentially allow the spectral moments to become sensitive to different aspects of production power spectra, it remains unclear how the frequency range is best selected. Frequency ranges up to around 10 kHz appear slightly more stable, since the tails of the resulting distributions appear to be slightly less variable with respect to an appropriate zero reference.

Figure 5.9 shows the 1st and 2nd spectral moments for all voiceless fricatives (in all vowel contexts), using a frequency range of 0–10 kHz, and a zero reference of −10dB SPL/Hz, as produced by all male and female subjects. This combination of frequency scale and zero reference was found to give the best spectral moments: the correlation coefficients are low ($C = -0.11$ and $C = -0.13$ for males and females respectively), and the separation between the sibilants is good ($J = 3.55$ and $J = 8.43$ for males and females respectively). Note that the non-sibilants remain completely overlapped ($J = 0.70$ and $J = 0.08$ for males and females respectively), due to their non Gaussian-like spectral shape. The apparent increased separation amongst female sibilants compared to the male sibilants, is a good example of how careful consideration must be taken before drawing conclusions from spectral moments. It may be tempting to draw some conclusion about 'better articulation' by the female subjects compared to their male counterparts. However, in light of what we now know of the properties of spectral moments, it seems highly likely that the greater separability results from *some increase in the suitability of female spectra for characterisation by spectral moments.*

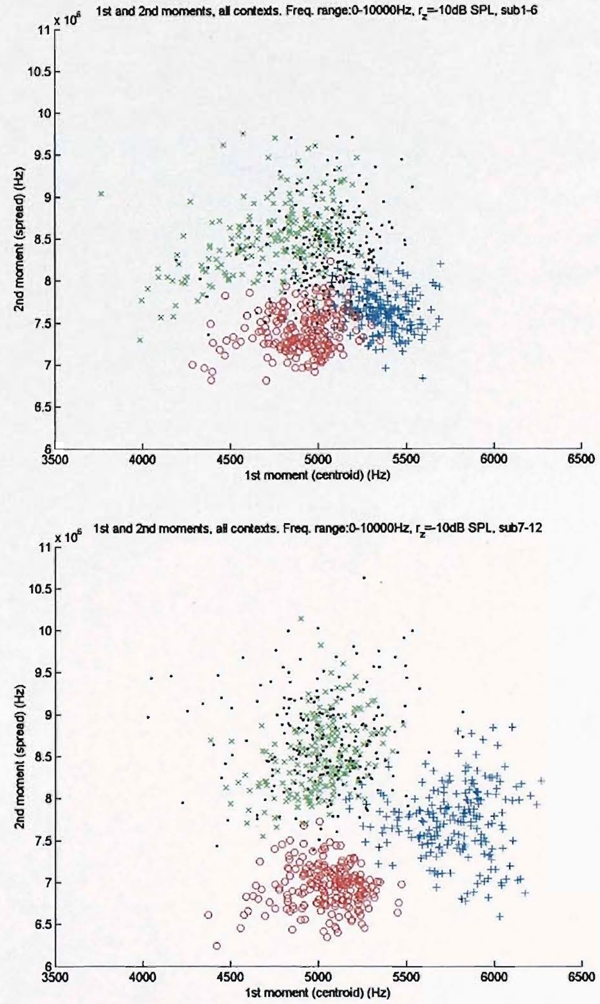


FIGURE 5.9: Best spectral moments for all fricatives in all contexts, all male subjects (top) and all female subjects (bottom). 0–10-kHz frequency range multitaper spectra, $r_z$=-10dB SPL. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

## 5.2 Variations of spectral moments over time

In order to gain an understanding of the typical range of expected variations *within* a fricative production, the spectral moments can be plotted over time.

Plots of spectral moments over time have been generated for all subjects and for all voiceless fricatives in all vowel contexts. What follows is generally true of the typical characteristics observed. To aid our discussion, we make use of a small set of 'typical' spectral moments over time.

The moments over time of /iFi/ segments (where /F/ is one of the voiceless fricatives) from productions of subject M-01, are shown in figures 5.10 to 5.13, using a frequency range of 0 to 12 kHz, and $r_z$=0dB SPL/Hz. This combination of frequency range and zero reference have been chosen to maximise the sensitivity of the spectral moments, while limiting the number of 'spurious' spectral moment values.

The vowel-fricative and fricative-vowel 'boundaries' are reasonably straightforward to recognise from these plots: within the 'central' vowel and 'central' fricative regions, the plots tend to change slowly over time; however, at boundaries, the moments often 'jump', or 'dip' in a pronounced manner. A good example of this can be seen in the central bottom plot of Figure 5.10, where the [ʃ-i] boundary produces a characteristic jump in all the spectral moment values. Often the boundaries are not so pronounced, but invariably some distinct wiggle can be observed. At the boundaries of the fricative, it is common for the overall amplitude of the speech (and the entire spectrum) to drop. Often this will result in a spectral distribution that drops below the zero reference at some frequencies, causing the spectral moments to become highly unstable, and resulting in the jumps seen at the fricative boundaries of the spectral moment plots over time.

The shapes of the even order moment curves over time seem to be insensitive to fricative place. This may at least partially be a result of the insensitivity of the spectral moments due to the distributions not approaching zero at the tails.

The odd-order moments seem to be more sensitive, both to the place of production, and to changes that occur over time. Productions of /s/ and /ʃ/ tend to have smooth first spectral moment plots over time, steadily rising after onset, and falling just before offset. The third spectral moments for /s/ and /ʃ/ follow a very similar trajectory, but mirrored horizontally, so that it has an appearance similar to the first moment, but 'upside down'.

The non-sibilant spectral moment shapes over time are not greatly different from the sibilants. However, the odd moments are generally subject to a higher amount of variability, notably so over the central portion of the fricative. Considering the definition of a stationary process is one whose statistical characteristics are independent of time (see §2.2.2.2), it is reasonable to say that the non-sibilants often appear non-stationary, even over their central regions. This has important implications for the use of time-averaging when trying to generate a consistent spectral estimate of mid-non-sibilant spectrum.

Spectral changes over time clearly occur, most especially during non-sibilant production. It has previously been expected that the characteristics of such changes over time may be related to the place of production for non-sibilants. However, while the spectral moments certainly seem to suggest such changes, characteristic differences in these spectral moment changes over time are not apparent.

## 5.3 Discussion

It has been shown that the use of multitaper spectra in the calculation of spectral moments does not produce greatly differing results from those calculated using modified periodogram spectra, with the exception of stabilising the higher-order moments due to reduced estimate variance in the tails, where the spectrum is especially prone to dropping below the zero reference.

Since the overall variance of the estimate is greatly reduced, the zero-reference can be raised to a higher level without risking much of the spectrum becoming 'negative': if the zero-reference can be raised sufficiently, the spectrum becomes better-suited to being modelled by a normal curve, and hence, more sensitive to changes in spectral shape.

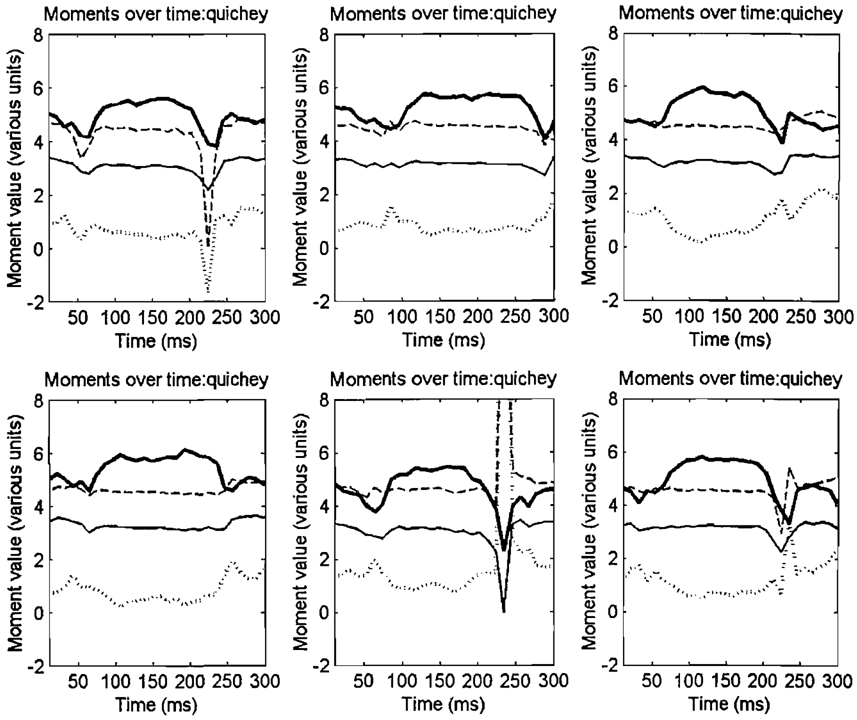


FIGURE 5.10: 1st (solid thick line), 2nd (solid thin line), 3rd (dotted line), and 4th (dashed line) moments from multitaper spectra over time of 6 tokens of [iʃi] from "quichey" by subject M-01.

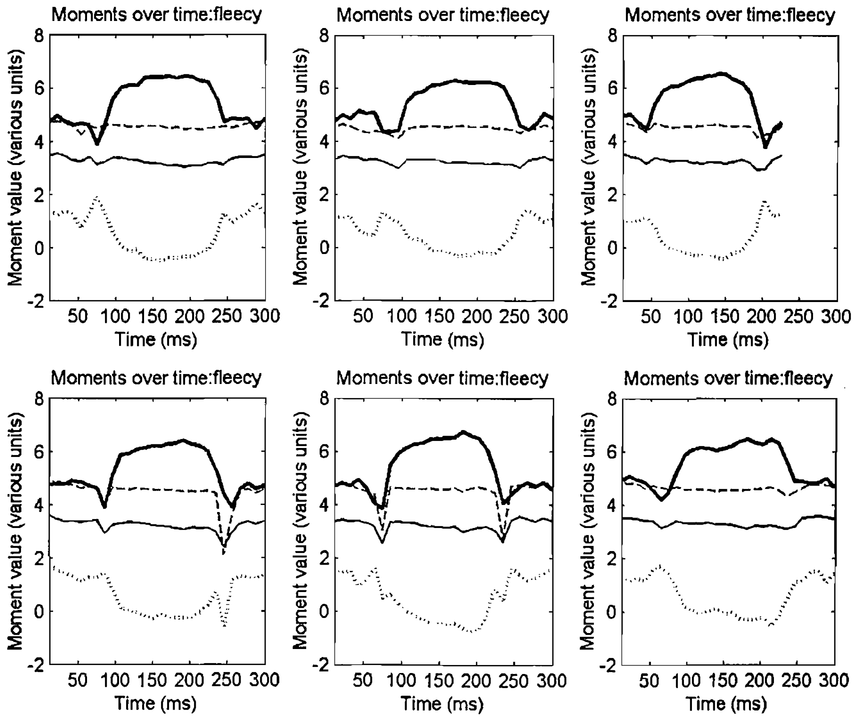


FIGURE 5.11: 1st (solid thick line), 2nd (solid thin line), 3rd (dotted line), and 4th (dashed line) moments from multitaper spectra over time of 6 tokens of [isi] from "fleecy" by subject M-01.

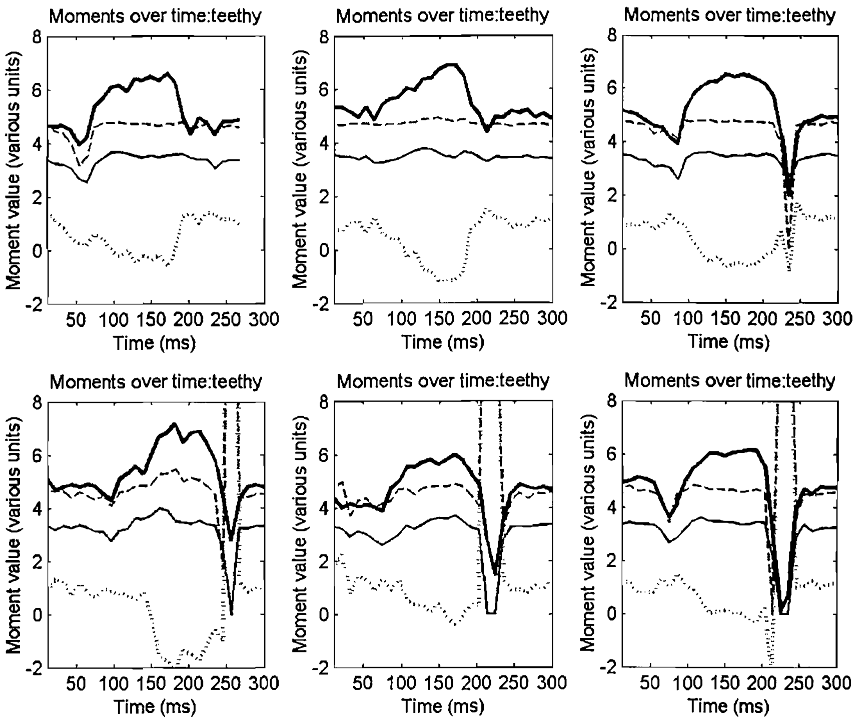


FIGURE 5.12: 1st (solid thick line), 2nd (solid thin line), 3rd (dotted line), and 4th (dashed line) moments from multitaper spectra over time of 6 tokens of [iθi] from "teethy" by subject M-01.

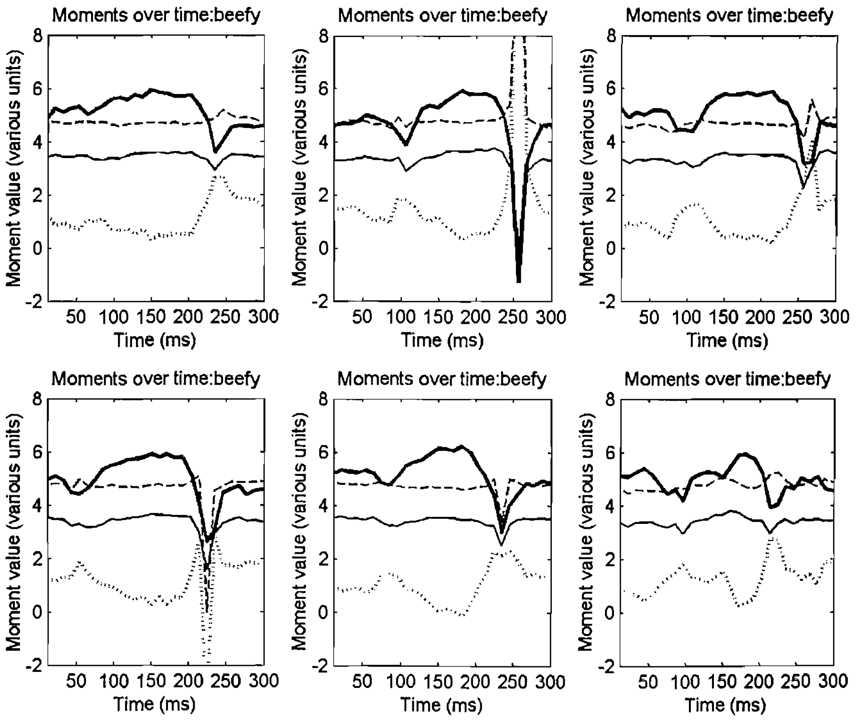


FIGURE 5.13: 1st (solid thick line), 2nd (solid thin line), 3rd (dotted line), and 4th (dashed line) moments from multitaper spectra over time of 6 tokens of [ifi] from "beefy" by subject M-01.

Figure 5.7 demonstrates the high degree of correlation amongst odd, and even moments. This is at least partially a result of the inescapable high-tailed spectral distributions that are being modelled as Gaussian curves. These correlation patterns cannot be greatly altered by using different frequency ranges, or zero references, due to the range of production variations.

The best results for spectral moments are therefore elicited by the 1st and 2nd moments. Spectral moments for fricative sibilant spectra have been shown to have good separation, if an appropriate combination of frequency range and zero reference is used, as seen in Figure 5.9.

A few existing studies have attempted several different approaches to setting the zero reference, although in some cases, the issue is not addressed at all. Functions of the maximum peak amplitude have been used to determine the zero reference. However, this approach will invariably introduce nonlinear correlations between the moments. More significantly, such nonlinear approaches make the interpretation of results very difficult. Such approaches may have applications in automatic speech recognition tasks, but they are of limited value if the results are to be used in some analysis of speech production.

The variations in spectral moments over time during the production of fricatives has been examined. No distinguishing characteristics of the variations over time have been found. However, strong evidence suggesting the non-stationarity of the voiceless fricatives has been presented. Notably, the non-sibilant voiceless fricatives appear to be particularly variable over time. The implications for these findings are twofold.

Firstly, since the moments of the process are known to vary over time, the validity of using time-averaging methods for improving the spectral estimate is called into question. The quite distinct non-stationary odd moments of the non-sibilants suggest that the use of time-averaging techniques will not necessarily produce useful results.

Secondly, while no features of the variation over time of the spectral moments have been found to distinguish the non-sibilants, the evidence of their non-stationarity strongly supports the notion that such temporal distinguishing cues exist. In order to catch such temporal changes, it may be necessary to more carefully capture and track distinct spectral features, rather than just the broad distribution of energy.

# Chapter 6

# Measurement, and characteristics of variation

Until now, a very limited amount of work has been undertaken to attempt to capture variations that occur across spectra. When it has, it has tended to use very broad descriptive techniques, that relay little information about detailed spectral features. This has largely been due to the lack of a suitable spectral estimation tool that both produces spectral estimates with small variance error, and also that does not need to compound data over which the analysis of variance is under examination, such as time and ensemble averaging techniques.

In this chapter, we undertake such analyses, using spectral estimates calculated using the multitaper methodology. We have seen that multitaper analysis excels when short windows of stochastic time-series data are to be analysed, and an estimate with low variance error is required. Such an analysis tool makes estimation of the variance of the *process* possible.

We begin by exploring the typical spectral variability that can occur in productions of voiceless fricatives across speakers of the same gender. This analysis allows us to view the maximum variability we are likely to encounter when we later consider the productions, for example, within a single speaker's fricative tokens.

Analysis of the spectral variation of the sibilants is found to produce some very useful results. We then turn our attention to the non-sibilants, for which these new analysis of spectral variance methods provide interesting new insight.

Appendices C, D, E and F contain plots relating to the results of this chapter.

## 6.1 Spectral variability in voiceless fricative production

Of interest is an estimate of the total variability within the productions of a single fricative, by speakers of the same gender. That is, for a given fricative, how much spectral variation exists within the speakers of a single gender, given that vowel context, and precise time position within the production are unknown?

### 6.1.1 Analysis of within-gender spectral variance

We begin by limiting our analysis to a 64-ms data window in the centre of all recorded fricative tokens. The central portion of each fricative token has deliberately been considered since it should be the region least influenced by vowel context. Clearly, as the window of analysis extends towards the fricative boundaries, an increase in variability can be expected. For now, the task is restricted to the fricative centre.

In order to ensure that variability present during the central fricative region is captured, the 64-ms window is divided into six adjacent 10.6-ms windows. These windows are treated independently for the purposes of variability measurement.

The multitaper 10.6-ms mid-fricative /F/ spectrum on decibel scale is denoted

$$\left(\hat{\Omega}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r,n]} \tag{6.1}$$

where $g$ is the subject gender, $s$ is the subject number and $1 \leq v \leq 6$ corresponds to the six vowel contexts /uFi/, /iFi/, /əFi/, /iFə/, /uFə/ and /əFu/; $1 \leq r \leq 6$ is the repetition number, and $1 \leq n \leq 6$ is the particular 10.6-ms window within the token. For the time being, the multitaper spectral estimates are calculated using 512-point data windows.

The sample production mean spectrum of fricative /F/ across all male (or female) subjects, in all vowel contexts, is then given by

$$\begin{aligned} \mu_{[g,/\mathrm{F}/]} &= \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[g,/\mathrm{F}/]} & (6.2)\\ &= \frac{1}{1296}\sum_{v=1}^{6}\sum_{s=1}^{6}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\hat{\Omega}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r,n]}, & (6.3) \end{aligned}$$

and the sample production spectral variance under the same criteria is given by

$$\begin{aligned} \varsigma^2_{[g,/\mathrm{F}/]} &= \mathrm{var}\left\{\hat{\Omega}(f_k)\right\}_{[g,/\mathrm{F}/]} & (6.4)\\ &= \frac{1}{1295}\sum_{v=1}^{6}\sum_{s=1}^{6}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\left(\hat{\Omega}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r,n]} - \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[g,/\mathrm{F}/]}\right)^2. & (6.5) \end{aligned}$$

Higher order moments could be calculated, but for now the mean and variance shall suffice.

The results of sample production spectral mean and variance are shown in figures 6.1 (males) and 6.2 (females). The solid lines represent the mean spectrum $\mu_{[g,/\mathrm{F}/]}$, while the dashed lines show the spectral variance $\varsigma^2_{[g,/\mathrm{F}/]}$. This is the first time that the variance of production has been estimated: such an analysis of production variance using modified periodograms produces variance plots that are of no use due to the incorporation of the *estimate* variance.

These plots offer some exciting new insights. While the mean plots act as a guide to the position in the spectrum being examined, as well as showing general spectral 'features', the variance plots are perhaps of much greater interest. The spectral variance plots indicate at which frequencies the sound intensity is subject to high variability, and at which frequencies it is more consistent.

We begin by considering both male and female /s/ productions. In both plots, the mean spectrum

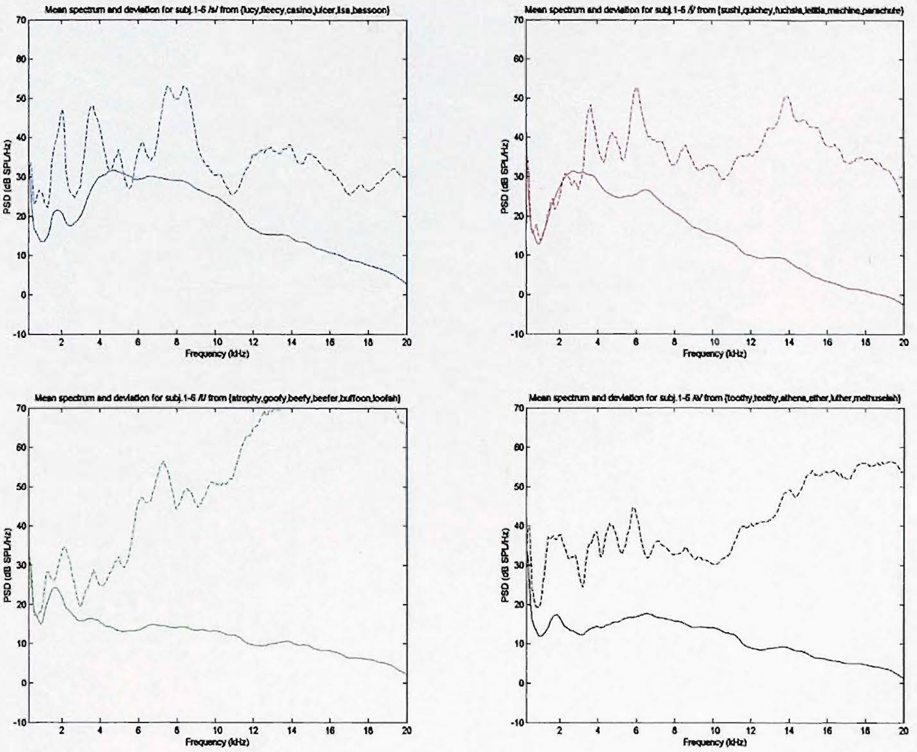


FIGURE 6.1: Mean and variance of central /s/ (top-left), /ʃ/ (top-right), /f/ (bottom-left) and /θ/ (bottom-right) spectra, taken from non-sibilants in all contexts, by all male subjects. (The maximum variance for /f/ reaches around 75dB SPL/Hz, and so is just out of view in this plot.)

suggests a low peak generally appears at ∼2 kHz, surrounded by troughs at ∼1 kHz and ∼3 kHz. The low peak appears to be the subject of a large amount of variability; however, the troughs surrounding it have particularly small variance in comparison to the rest of the spectrum.

Over the ∼3 to ∼5-kHz range, the mean /s/ spectra rise quite rapidly, and then remain high over the ∼5 to ∼10-kHz range for men, and the ∼5 to ∼13-kHz range for women. The initial rapid slope around the ∼3 to ∼5-kHz interval appears to be the subject of a high degree of variability. If the frequency positions of the lower and upper points of this slope are subject to a small amount of shifting left and right along the frequency axis, then this would account for this high degree of spectral variance in this region. The ∼5 to ∼13-kHz range is generally subject to a high degree of variance, with the exceptions of one or two regions of smaller variance, specifically ∼6 kHz for men and ∼8 kHz for women.

There is another dip in the variability of male /s/ productions around 11 kHz, although no similar dip is found for female productions. Above ∼11 kHz, both male and female spectra are subject to large variation, at least when considered across speakers and vowel contexts.

The mean /ʃ/ spectra for both males and females show a pronounced trough at ∼1 kHz, followed by a steep slope up to the main peak, which appears at around 3 kHz for men and 4 kHz for females. Again, the position of the slope seems to be subject to a high degree of variability for the females, but much less so for the male productions. In both cases however, the spectral peak coincides with a region of low variance, and this is of great interest. Male /ʃ/ production spectra

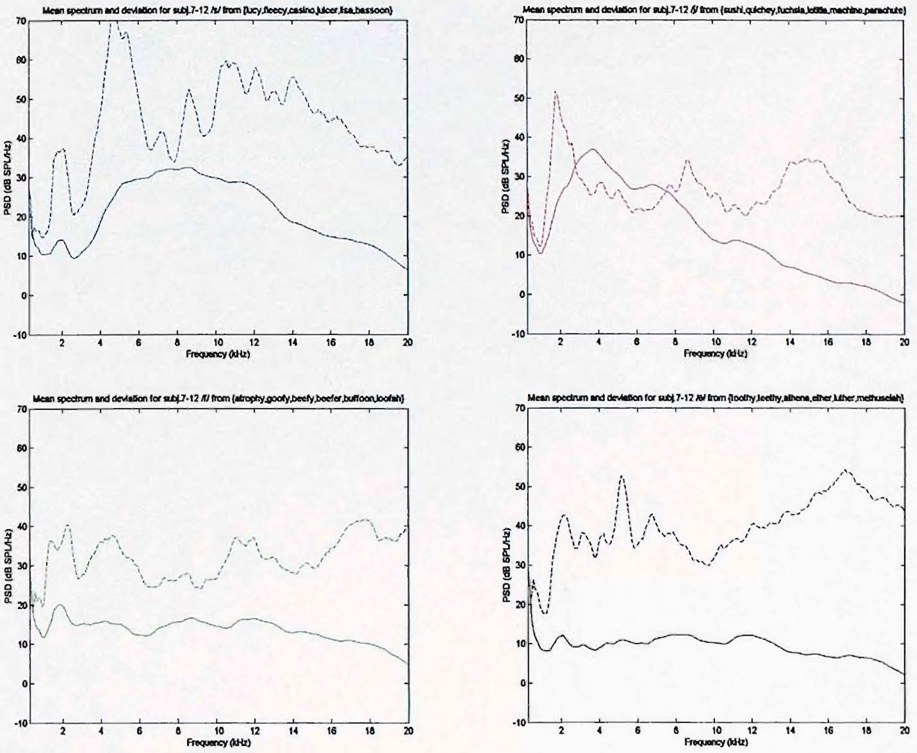


FIGURE 6.2: Mean and variance of central /s/ (top-left), /ʃ/ (top-right), /f/ (bottom-left) and /θ/ (bottom-right) spectra, taken from non-sibilants in all contexts, by all female subjects.

are generally the subject of large variance at frequencies above the main mean spectral peak, although for females this variability is reduced at these higher frequencies.

While the mean sibilant spectra are rather distinct in terms of their main peak and troughs, the non-sibilant mean spectra appear very similar to each other. Both have mean peaks around 2 kHz, and slight troughs either side: just below 1 kHz, and 3 kHz (although for the mean female spectra, the 3-kHz trough is hardly pronounced at all). Of interest is that the troughs coincide with low production variance, except perhaps for female /θ/ spectra. Otherwise, the mean spectra are generally featureless, and tend to have very high production variance over the majority of the frequency range. For males, the /f/ spectra contain the most production variance above, say, 6 kHz; however, for female non-sibilants, /θ/ contains the most spectral variance above, say, 2 kHz. It appears the most striking aspect of the non-sibilants is their lack of spectral features, and large production variability. We shall return to the non-sibilants in §6.1.3.

### 6.1.2 Characteristics of the voiceless sibilants

The frequencies at which mean spectral features of the sibilants occur often coincide with frequencies of low production spectral variance. In particular, it can be seen that the spectral maxima for /ʃ/ productions, approximately coincide with the pronounced spectral trough of /s/ productions, and it is noted that in both cases, these mean spectral features are coupled with low production variance.

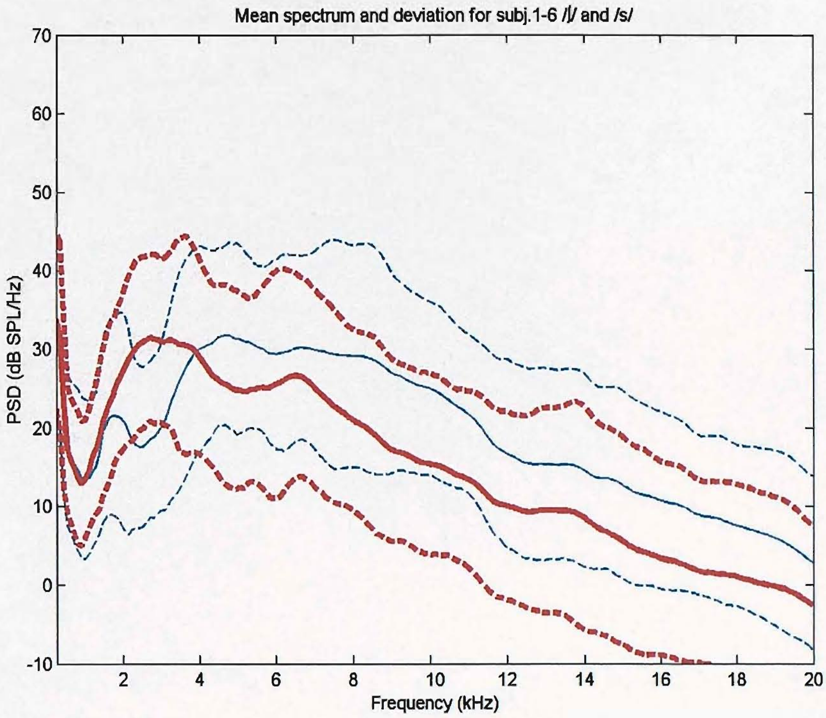


FIGURE 6.3: Sibilant tokens mean spectra (solid lines) with two standard deviations bounds (dashed lines) either side. /s/ tokens (blue, thin) and /ʃ/ tokens (red, thick) in all contexts, from all male subjects (see text).

In order to view to what extent these spectral features of low variability are distinguishing features of the sibilant, it is straightforward to generate plots of the mean spectra, with bounds of two standard deviations above and below the mean spectra. Such plots of male /s/ and /ʃ/ spectral 'variations' are shown superimposed in Figure 6.3. The plot shows $\mu_{[\text{male},/s/]}$ (solid thin blue line), $\mu_{[\text{male},/\int/]}$ (solid thick red line), $\mu_{[\text{male},/s/]} \pm 2\varsigma_{[\text{male},/s/]}$ (dashed thin blue lines), and $\mu_{[\text{male},/\int/]} \pm 2\varsigma_{[\text{male},/\int/]}$ (dashed thick red lines).

The region around 2.5 kHz is most striking: it shows that the maximum magnitudes for /ʃ/ productions reach similar values to the minimum magnitudes for /s/ productions. That is, the spectral trough at 2.5 kHz of /ʃ/ productions rarely rises above 27dB SPL/Hz, while the spectral peak at the same frequency in /s/ productions rarely drops below 22dB SPL/Hz.

The frequency $f_\xi$ of the main spectral peak in a token sibilant spectrum is located

$$\hat{\Omega}(f_\xi)_{[/\mathrm{F}/,v,s,r]} = \max_{f_k > 1\mathrm{kHz}} \left\{ \left(\bar{\hat{\Omega}}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r]} \right\} \tag{6.6}$$

where[1]

$$\left(\bar{\hat{\Omega}}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r]} = \frac{1}{6}\sum_{n=1}^{6} \left(\hat{\Omega}(f_k)\right)_{[g,/\mathrm{F}/,v,s,r,n]}. \tag{6.7}$$

[1] The time averaged spectrum is used for clarity purposes only, so that a single point on the scatter plot corresponds to a single token. The scatter points using every window are very similar, but of course, much denser.

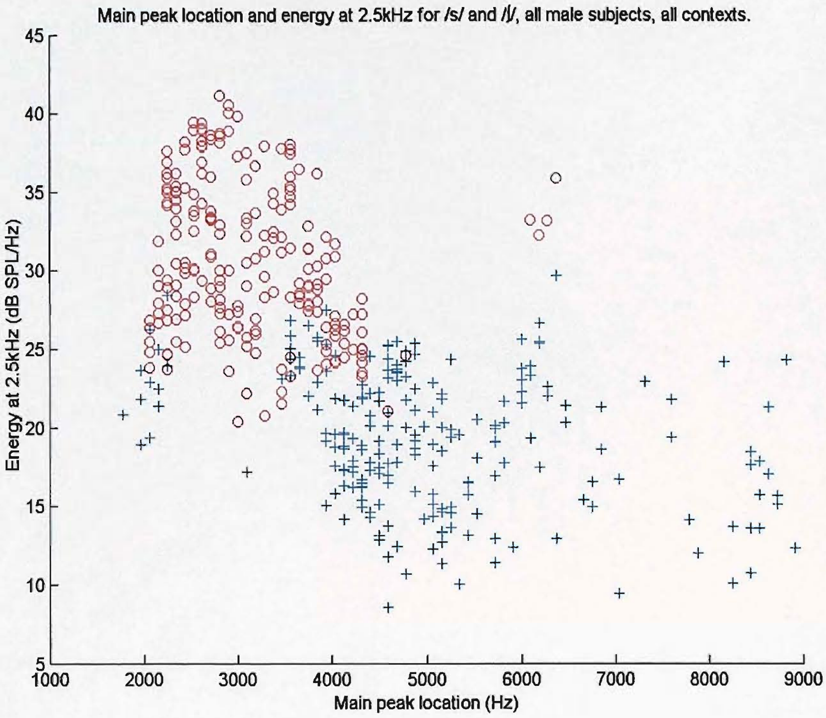


FIGURE 6.4: Scatter plot of energy in 2.5-kHz interval, against peak frequency for all male sibilant tokens. Blue plus-signs are /s/, while red circles are /ʃ/.

A scatter plot of the energy in the 2.5-kHz band (2484 Hz–2578 Hz) against spectral peak locations is shown in Figure 6.4 for all male sibilant tokens. This plot shows clearly the power at 2.5 kHz as a function of peak frequency for the sibilants. Both measures seem to be heavily influenced by place, although there is some overlap.

In fact, the two sibilants are distinguished well using only the energy measure at 2.5 kHz. We might expect that this energy measure will also be influenced by the total intensity of the fricative. A scatter plot of energy at 2.5 kHz against total spectral energy is given in Figure 6.5. This time, the correlation between the total fricative intensity, and the energy in the 2.5-kHz band can be recognised ($C = 0.54$ and $C = 0.66$ for /s/ and /ʃ/ respectively), and this aids the distinction between the two sibilants ($J = 4.53$).

We now perform similar analysis on the female sibilant tokens. We begin with Figure 6.6, showing $\mu_{[\text{female},/s/]}$ (solid thin blue line), $\mu_{[\text{female},/ʃ/]}$ (solid thick red line), $\mu_{[\text{female},/s/]} \pm 2\varsigma_{[\text{female},/s/]}$ (dashed thin blue lines), and $\mu_{[\text{female},/ʃ/]} \pm 2\varsigma_{[\text{female},/ʃ/]}$ (dashed thick red lines). This time, the region of maximal separation is shifted up in frequency slightly, to around 3 kHz. It can clearly be seen that in this region, the two standard deviation confidence bounds for the sibilant spectra do not overlap at all, suggesting a distinguishing feature of the sibilants. The scatter plot of spectral peak frequency against energy in the 2953-Hz–3047-Hz frequency interval in Figure 6.7, demonstrates how distinguishing the energy measure is, even compared to a measure such as the main peak frequency.

Figure 6.8 highlights the strength of the energy measure in a narrow band around 3 kHz, as

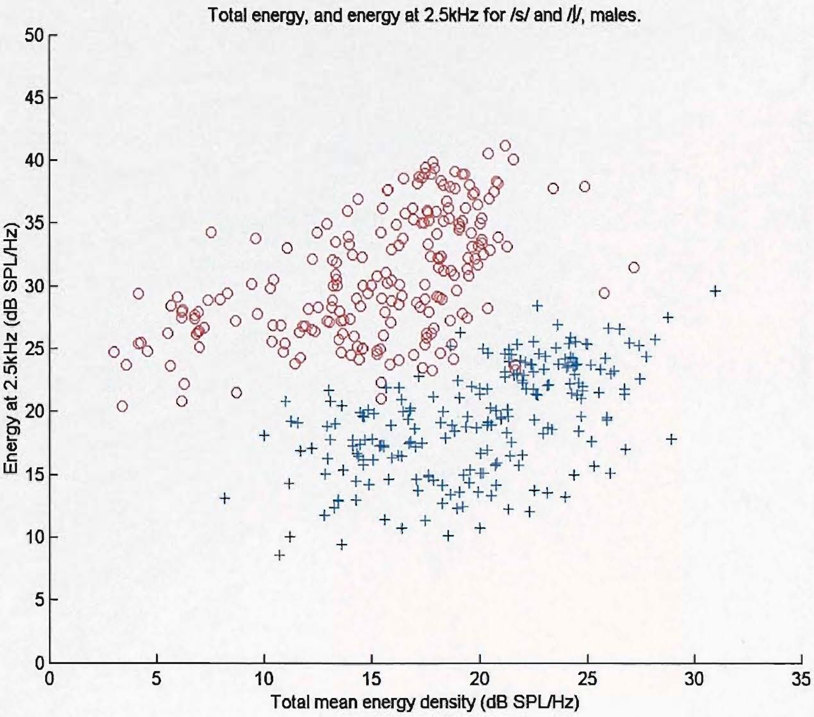


FIGURE 6.5: Scatter plot of energy in 2.5-kHz interval, against total spectral intensity for all male sibilant tokens. Blue plus-signs are /s/, while red circles are /ʃ/.

a distinguishing feature of the sibilants, by plotting it against the total spectral energy. The sibilants are completely separated using these energy measures for female tokens ($J = 18.23$), and the corellation of overall amplitude to energy density around 3 kHz can be seen ($C = 0.31$ and $C = 0.61$ for /s/ and /ʃ/ respectively).

### 6.1.3 Variability in the non-sibilants

As was shown in figures 6.1 and 6.2, analysis of the mean and variance of non-sibilant spectra reveal neither distinguishing spectral features, nor prominent regions of low variance. Superimposed plots of spectral mean with two standard deviation bounds can be produced as before. Figure 6.9 shows $\mu_{[g,/f/]}$ (solid thick green line), $\mu_{[g,/\theta/]}$ (solid thin black line), $\mu_{[g,/f/]} \pm 2\varsigma_{[g,/f/]}$ (dashed thick green lines), and $\mu_{[g,/\theta/]} \pm 2\varsigma_{[g,/\theta/]}$ (dashed thin black lines), for $g$=males (top plot) and $g$=females (bottom plot). This plot clearly demonstrates an important factor in confounding the attempts to classify the non-sibilants by spectral shape: the mean shapes are very similar, and have such high variability that the two-standard deviation confidence interval overlaps everywhere.

Both spectra have a common trough around 1 kHz, and a main peak located around 2 kHz, although these features are subject to a large variance. Generally speaking, /θ/ tokens appear to have lower amplitude than /f/ tokens, but again, there is a large degree of variability, so these points are in no way distinguishing features.

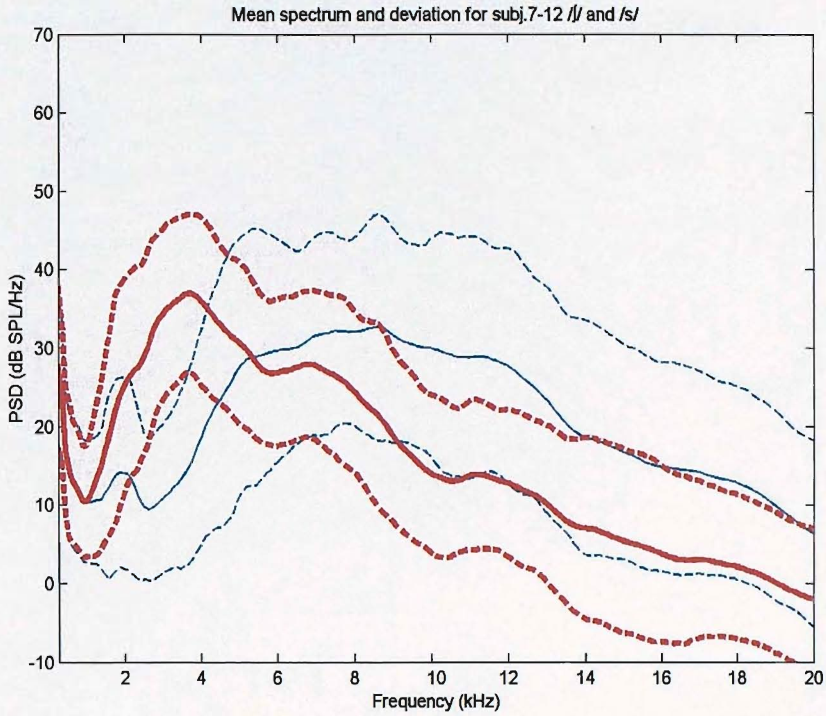


FIGURE 6.6: Sibilant tokens mean spectra (solid lines) with two standard deviations bounds (dashed lines) either side. /s/ tokens (blue, thin) and /ʃ/ tokens (red, thick) in all contexts, from all female subjects (see text).

A plot of spectral peak position and total spectral energy for mid-fricative male tokens, for all voiceless fricatives is shown in Figure 6.10. While no differences in distributions for the non-sibilants were expected in this plot, it does appear that trends emerge with peak position. It would appear that /f/ tokens often have a spectral peak around 2 kHz. On the occasions when the peak is not located here, it appears most likely to occur around 4 kHz, or alternatively 7 kHz. Very few other positions of spectral peak location are observed for mid-fricative /f/ tokens.

These trends are slightly contrasted with those of the mid-fricative /θ/ tokens. Spectral peaks of central /θ/ tokens also commonly occur around 2 kHz, although exceptions are far more common. The exceptions appear to be more evenly distributed over the 3–12-kHz range, although the region around 6 kHz is also well populated.

The equivalent plot for all female tokens is shown in Figure 6.11. The trends for female non-sibilant peak position are similar to those noted for the male tokens. The vast majority of mid-fricative /f/ tokens tend to have a peak around 2 kHz (more often than for the male productions). When not near 2 kHz, the peak tends to occur around 4.5 kHz, 8 kHz or 12 kHz, although these are quite rare. Mid-fricative /θ/ tokens also often have peaks near 2 kHz, but much less commonly than for /f/ tokens. The main peak position for the remainder of the /θ/ tokens has a more even likelihood distribution across the 3–18-kHz range, although there are significant clusters that overlap with the /f/ tokens than for male tokens.

For completeness, the plots of total mean energy density and energy density around 2.5 kHz for all male fricatives and around 3 kHz for all female fricatives are given in Figure 6.12. As we have

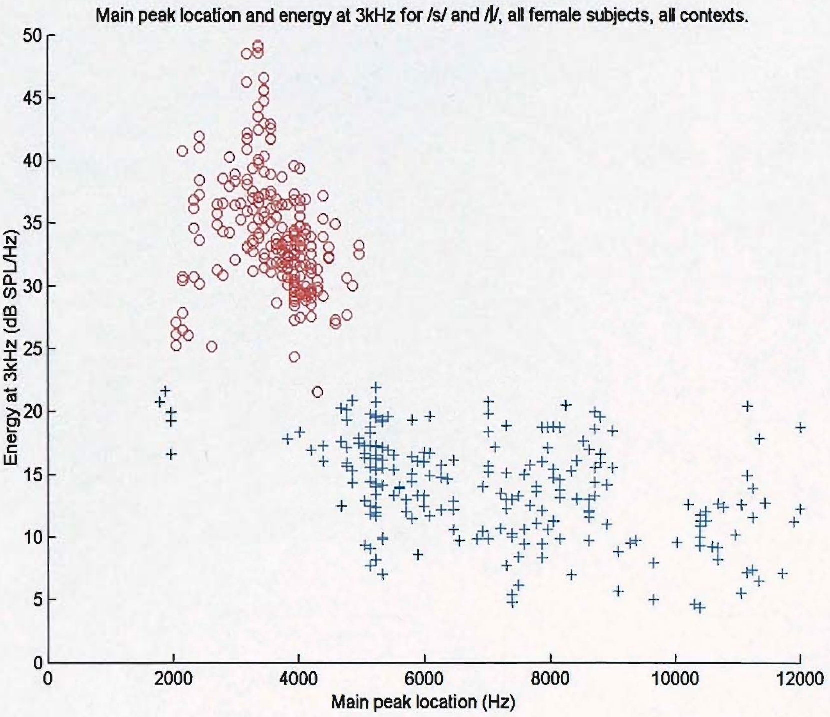


FIGURE 6.7: Scatter plot of energy in 3-kHz interval, against peak frequency for all female sibilant tokens. Blue plus-signs are /s/, while red circles are /ʃ/.

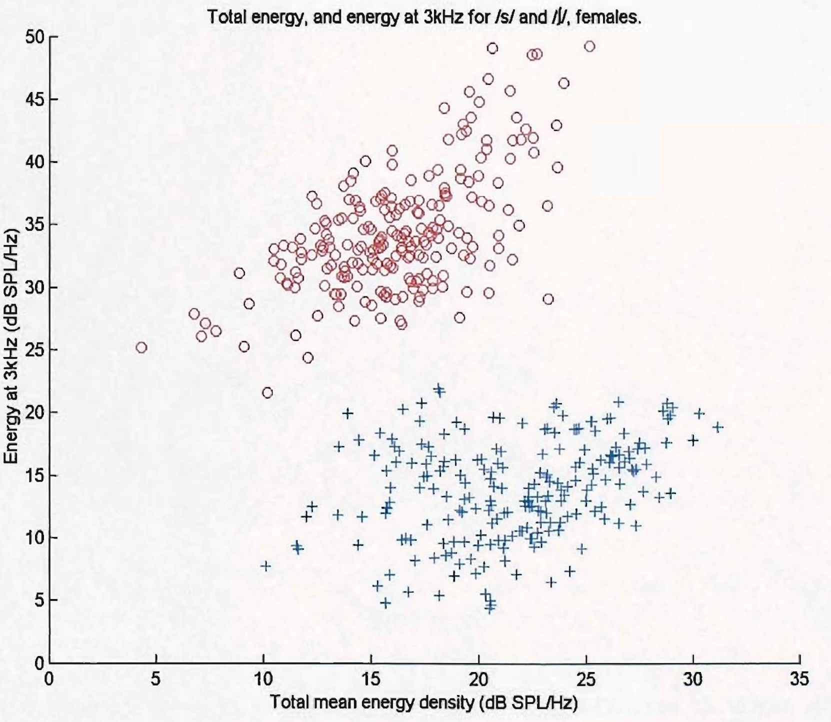
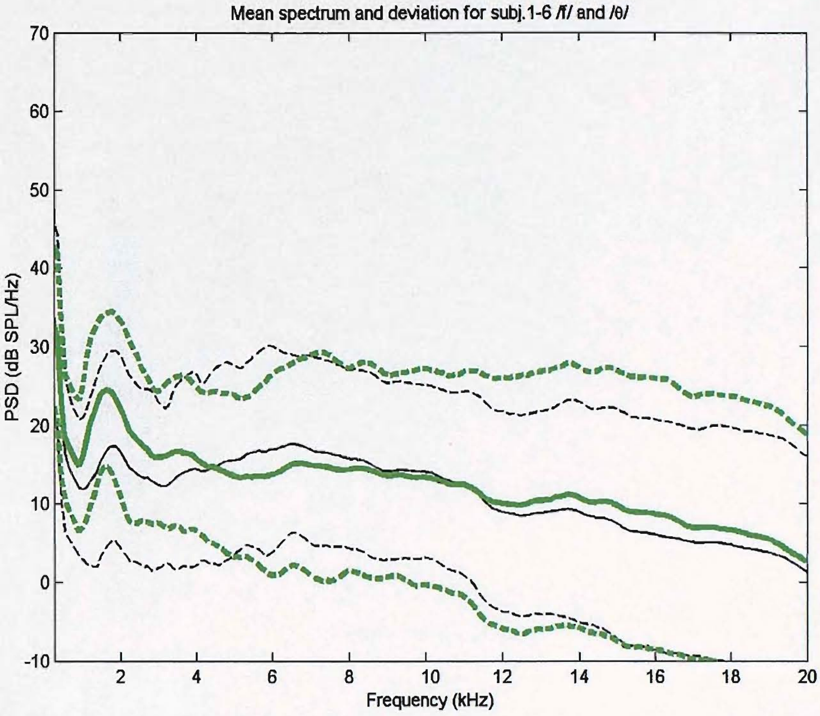


FIGURE 6.8: Scatter plot of energy in 3-kHz interval, against total spectral intensity for all male sibilant tokens. Blue plus-signs are /s/, while red circles are /ʃ/.



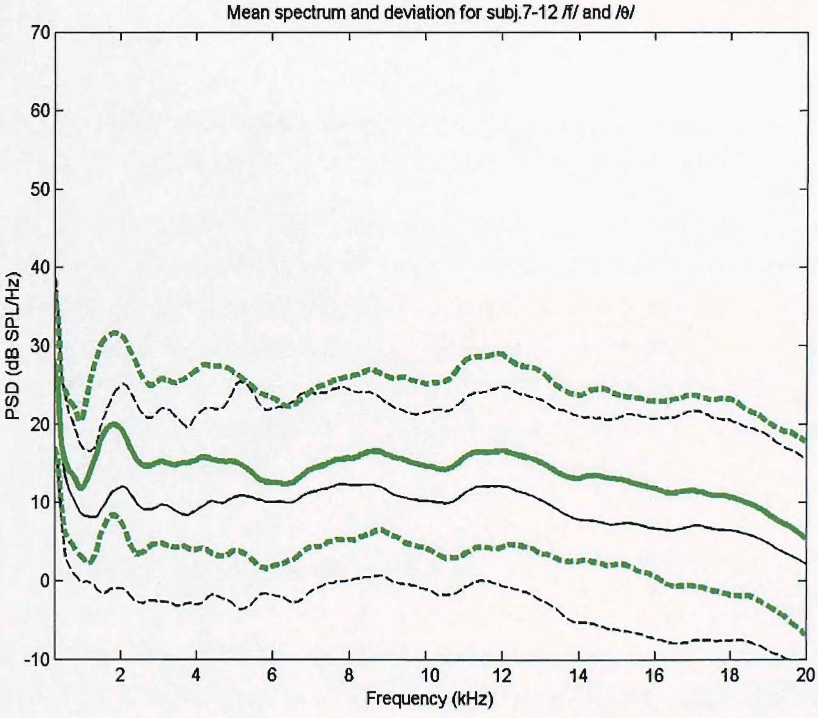


FIGURE 6.9: Non-sibilant tokens mean spectra (solid lines) with two standard deviations bounds (dashed lines) either side. /f/ tokens (green, thick) and /θ/ tokens (black, thin) in all contexts, from all male (top) and female (bottom) subjects.

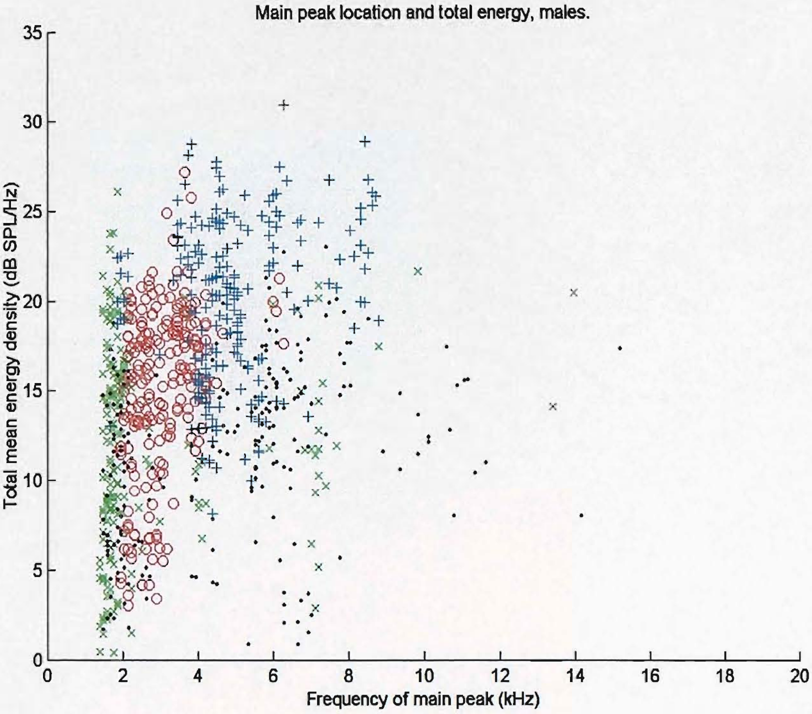


FIGURE 6.10: Spectral peak position and total spectral energy for mid-fricative male tokens, all voiceless fricatives. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

seen, such plots are particularly useful for separating the sibilants. Their use in separating the non-sibilants is not so obvious, although the distributions are slightly different.

Clearly, in order to determine the major differences in non-sibilant production, a more thorough analysis of the tokens must be performed. The large production variances observed for the non-sibilants may be the result of large differences across different speakers. It may additionally be the result of a large degree of production variation across fricative tokens in different vowel contexts for each speaker. Finally, the production variation could occur within each token. The total production variance seen in figures 6.1 and 6.2 may be due to any or a combination, of these potential sources of variation.

### 6.1.4 Within-speaker spectral variability

We begin by observing the non-sibilant token production variation by speaker. The sample production mean spectrum of fricative /F/ in all vowel contexts, for a given speaker $s$, is given by

$$\mu_{[/\mathrm{F}/,s]} = \overline{\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,s]}} \tag{6.8}$$

$$= \frac{1}{216}\sum_{v=1}^{6}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]}, \tag{6.9}$$

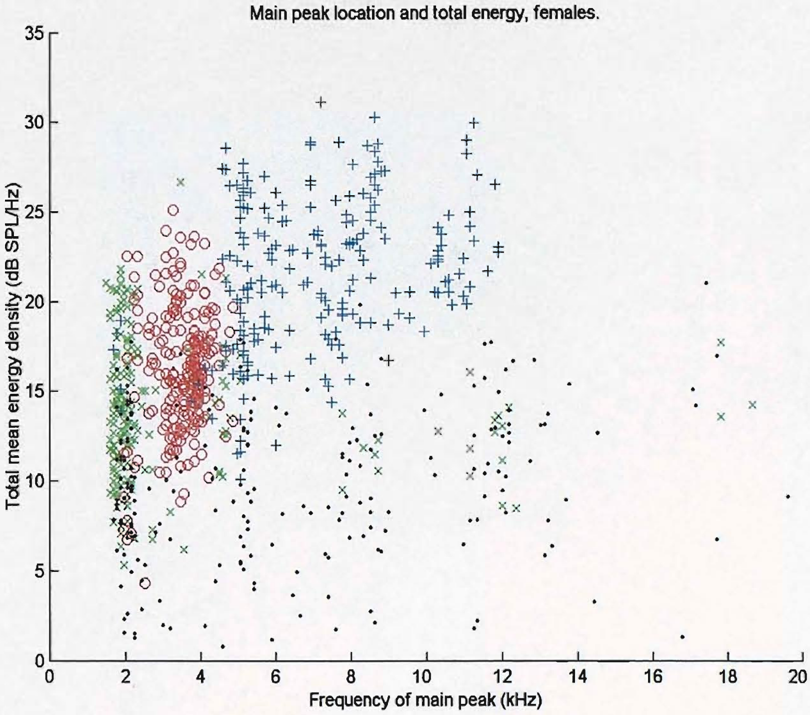


FIGURE 6.11: Spectral peak position and total spectral energy for mid-fricative female tokens, all voiceless fricatives. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

and the sample production spectral variance under the same criteria is

$$\varsigma^2_{[/\mathrm{F}/,s]} = \mathrm{var}\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,s]} \tag{6.10}$$

$$= \frac{1}{215}\sum_{v=1}^{6}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]} - \overline{\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,s]}}\right)^2 . \tag{6.11}$$

Figures C.9 to C.11 show the production spectral mean $\mu_{[/F/,s]}$ and variance $\varsigma^2_{[/F/,s]}$ of /f/ tokens in all vowel contexts, by male speaker. These can be compared with the /θ/ token production characteristics shown in figures C.12 to C.14.

Firstly, the mean spectra still appear to be very similar for /f/ and /θ/. Secondly, it can be seen that the production variance is still often high, although in some places it drops; most noticeably, the variance tends to be lowest around the 4-kHz region for both non-sibilants. The variance tends also to be high above around 12 kHz, although there are a few exceptions. Generally speaking, it appears that /f/ productions have a slightly lower overall spectral variability.

The male data can be compared to the female data in figures C.23 to C.28. Again, the non-sibilant mean spectra are very similar for /f/ and /θ/. The female production variances are generally high everywhere. Again, broadly speaking it would appear that /f/ productions have a slightly lower degree of spectral variance overall.

From these non-sibilant production mean and variance spectra, it can be seen that a significant

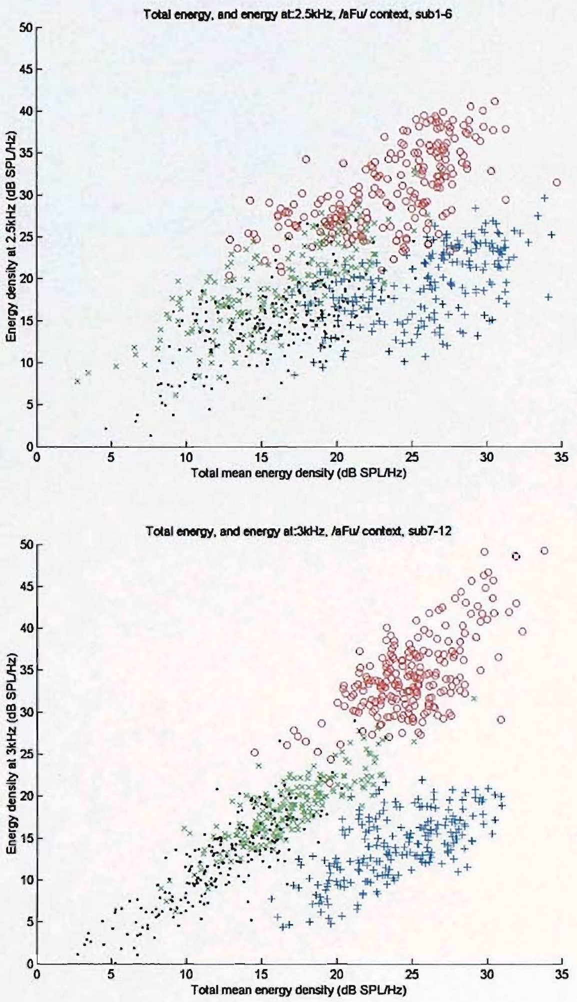


FIGURE 6.12: Total mean energy, and energy density around 2.5 kHz for males (top) and around 3 kHz for females (bottom) for all fricatives. Legend: [s] (blue plus-signs); [ʃ] (red circles); [f] (green crosses); [θ] (black dots).

proportion of production variability in the non-sibilants can be accounted for by a high degree of within-speaker variation. Some speakers certainly produce the non-sibilants more consistently than others however, and it remains to be seen whether the variability is a result of vowel context, across-token variations, or even due to within-token variations. The most general trend observed is that /f/ production appears to have slightly lower within-speaker variability than /θ/.

### 6.1.5 Within-vowel-context spectral variability

In order to estimate the mean spectrum for fricative /F/ in a given vowel context $v$ by a single speaker $s$, we use

$$\mu_{[/\mathrm{F}/,v,s]} = \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[/\mathrm{F}/,v,s]} \tag{6.12}$$

$$= \frac{1}{36}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]}, \tag{6.13}$$

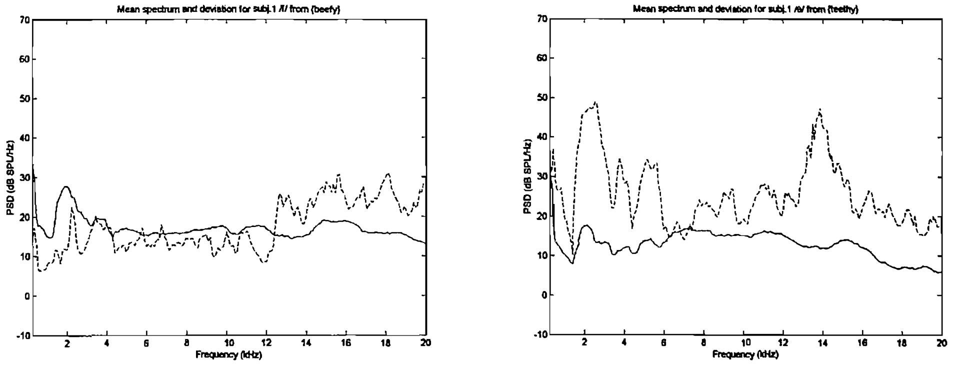


FIGURE 6.13: Mean spectrum (solid) and production variance (dashed) from central /f/ tokens in /ifi/ context (left), and /θ/ tokens in /iθi/ context (right), subject M-01.

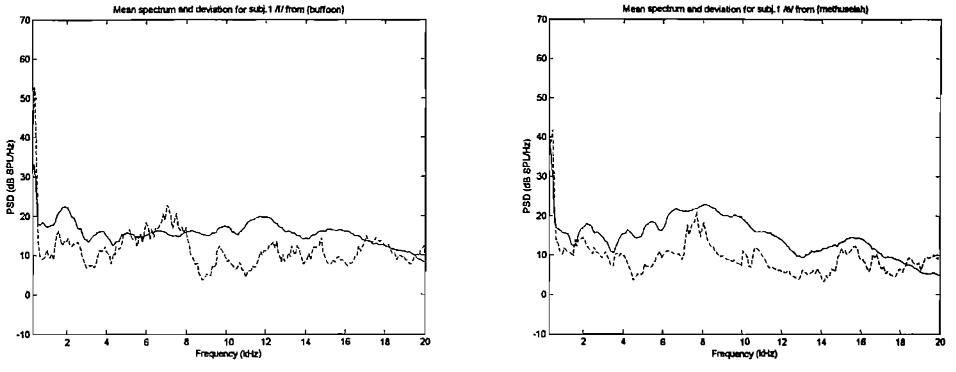


FIGURE 6.14: Mean spectrum and production variance (dashed) from central /f/ tokens in /əfu/ context (left), and /θ/ tokens in /əθu/ context (right), subject M-01.

the sample production spectral variance under the same criteria being

$$\varsigma^2_{[/\mathrm{F}/,v,s]} = \mathrm{var}\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,v,s]} \tag{6.14}$$

$$= \frac{1}{35}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]} - \overline{\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,v,s]}}\right)^2. \tag{6.15}$$

We are trying to ascertain whether the amount of variability seen for individual non-sibilant productions across vowel contexts, is due to changes in the production due to vowel context, or more inherent within the fricative production. On observation of $\mu_{[/F/,v,s]}$, and $\varsigma^2_{[/F/,v,s]}$ for /F/=/f,θ/ and $v$=/iFi,əFu/, it became apparent that a large amount of variability still exists within a particular vowel context.

As an example of this within-vowel-context variability, figures 6.13 and 6.14 show $\mu_{[/F/,v,s]}$ (solid lines), and $\varsigma^2_{[/F/,v,s]}$ (dashed lines), for /F/=/f,θ/, $v$=/iFi,əFu/, and $s$=M-01. Speaker M-01 has been selected for his apparently low production variability within the non-sibilants (see figures C.9 and C.12). However, it can clearly be seen that a high degree of spectral variation exists for /θ/ productions within the /iθi/ vowel context, suggesting that spectral variability for the non-sibilants is sometimes high, even within a given vowel context.

### 6.1.6 Within-token spectral variation

It is evident that distinguishing features of the non-sibilants are not present in the spectral shape. The mean spectra for the non-sibilants are extremely similar. In order to discover whether distinguishing features appear as variability over time, we can continue to calculate the *within-token* spectral variation, thus:

$$\begin{aligned}\mu_{[/\mathrm{F}/,v,s,r]} &= \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[/\mathrm{F}/,v,s,r]} && (6.16)\\ &= \frac{1}{N_w}\sum_{n=1}^{N_w}\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]}, && (6.17)\end{aligned}$$

the token's production spectral variance under the same criteria being

$$\begin{aligned}\varsigma^2_{[/\mathrm{F}/,v,s,r]} &= \mathrm{var}\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,v,s,r]} && (6.18)\\ &= \frac{1}{N_w-1}\sum_{n=1}^{N_w}\left(\left(\hat{\Omega}(f_k)\right)_{[/\mathrm{F}/,v,s,r,n]} - \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[/\mathrm{F}/,v,s,r]}\right)^2. && (6.19)\end{aligned}$$

Notice that the number of windows within the token has been changed from 6 to $N_w$. In order to achieve a satisfactory estimate of the spectral variability over the course of a fricative token, we need to try and maximise the number of sample data. In order to achieve this, the data window length was reduced from 10.7 ms (512-points) to 2.7 ms (128-points), but additionally, a larger portion of each token was considered (see §3.2.3). Since the data segments used in these calculations are of variable length, it is appropriate at this point to observe the segment lengths obtained from the calculations performed in §3.2.3. The mean lengths and the standard deviations of length are given in table 6.1

| (ms) | Mean length | Std. deviation |
|---|---|---|
| Male /s/ | 108.8 | 17.8 |
| Male /ʃ/ | 99.7 | 18.7 |
| Male /f/ | 112.1 | 26.1 |
| Male /θ/ | 100.3 | 24.9 |
| Female /s/ | 116.7 | 20.6 |
| Female /ʃ/ | 105.2 | 22.0 |
| Female /f/ | 130.4 | 25.5 |
| Female /θ/ | 111.8 | 30.5 |

TABLE 6.1: Means and standard deviations of fricative segment lengths calculated using the procedures described in §3.2.3. All units are milliseconds.

Example plots of within-token spectral variation are shown in Figure 6.15 and 6.16 for non-sibilant tokens produced by subject M-01. These plots are obviously smoother because of the reduced spectral resolution resulting from the reduced window sizes. These example plots are characteristic of all the male non-sibilant tokens, although exceptions are not rare. It appears that a reasonably distinguishing feature is that of the level of spectral variation within each non-sibilant token, /f/ tokens tending to be higher overall.

Also, note the higher variability in the lower frequencies below around 1 kHz. Since we are now

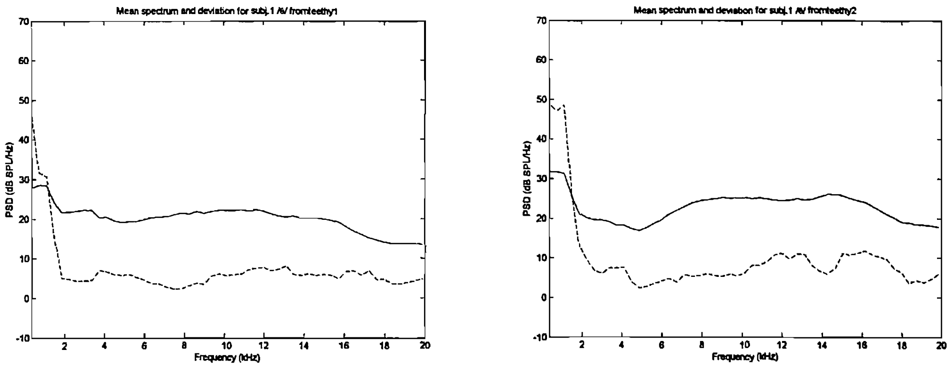


FIGURE 6.15: Example spectral means and variations for two individual /θ/ tokens in /iθi/ context, subject M-01.

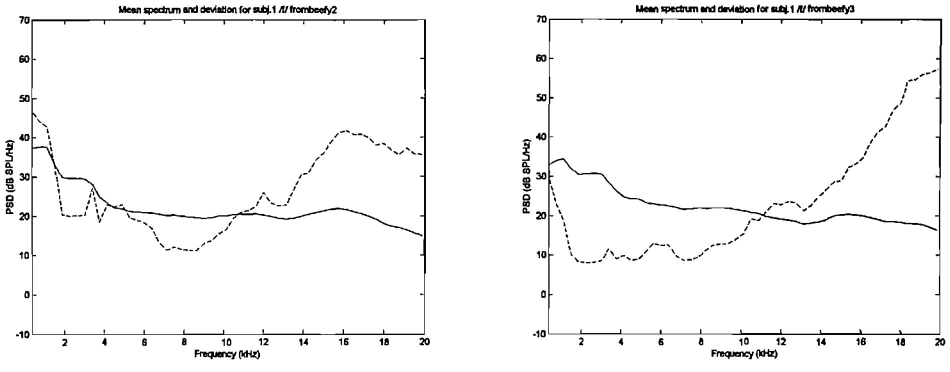


FIGURE 6.16: Example spectral means and variations for two individual /f/ tokens in /ifi/ context, subject M-01.

capturing a larger portion of the fricative token, we are including more data near the boundaries of the fricative. We may therefore expect this increase in variability at these lower frequencies.

Section F.3 in the Appendix shows additional plots of spectral variance over time for example fricative tokens by speaker and vowel context. These generally show that, while the sibilants have low spectral variance, the non-sibilants more frequently have a higher degree of spectral variance. The higher variability usually occurs within a wide frequency band of, say, 6 kHz in width; however, the location of this high-variance frequency band does not appear to have any consistent trends associated with it.

One approach to observing how distinguishing the level of within-token spectral variability is, is to plot the *total* spectral variability for each token for every token. A scatter plot of total spectral energy

$$\sum_k \overline{\left\{\hat{\Omega}(f_k)\right\}}_{[/\mathrm{F}/,v,s,r]} \tag{6.20}$$

against total spectral variability

$$\sum_k \operatorname{var}\left\{\hat{\Omega}(f_k)\right\}_{[/\mathrm{F}/,v,s,r]} \tag{6.21}$$

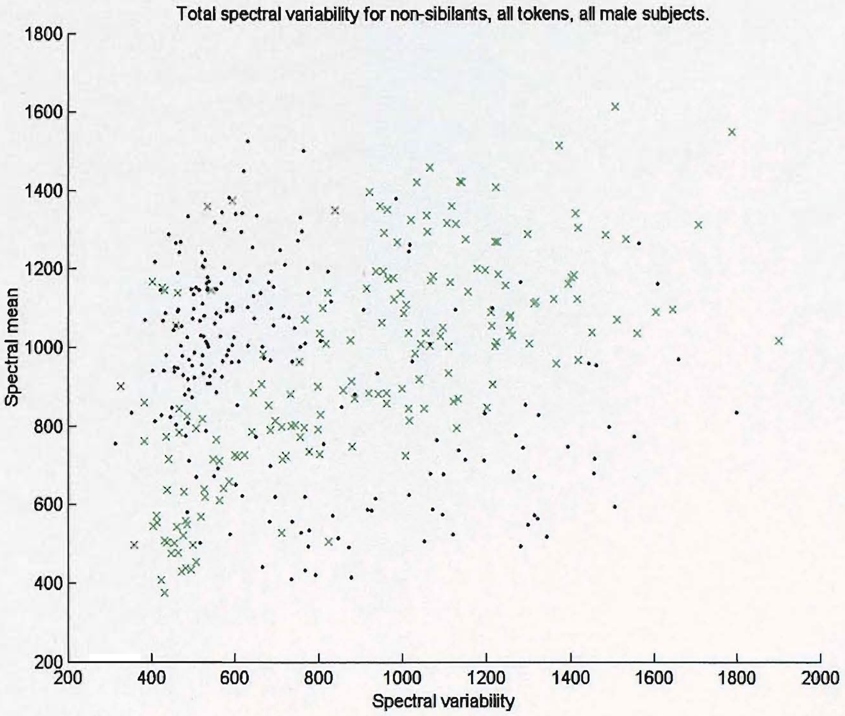


FIGURE 6.17: Scatter plot showing mean spectral energy against total spectral variance for all non-sibilant tokens, all male subjects. Green crosses are /f/ while black dots are /θ/.

is shown for all male non-sibilant tokens, in Figure 6.17. This plot indicates a possibly distinguishing characteristic of non-sibilant place ($p < 0.0001$, $J = 0.96$). The same plot for female non-sibilant tokens, shown in Figure 6.18 is not so convincing (despite $p < 0.0001$, $J = 0.96$). These results are not improved by separation by vowel context, Nevertheless there is certainly a strong suggestion that information pertaining to the production place is present within the spectral variability over the duration of the fricative.

Section F.4 in the Appendix separates these results by speaker and context (for /iFi/ and /əfu/ contexts). These results are of interest, since they suggest that vowel context does not consistently alter the degree of within-token variation, at least for the two vowel contexts considered.

### 6.1.7 Alternative measures of variability over time

A number of other methods have been used to attempt to capture the characteristics of the variation in spectrum over the course of non-sibilant tokens. Of these, one of the more successful methods is that of tracking the spectral peak through the non-sibilant. Observation of multitaper spectrograms suggests evidence that /θ/ spectra are generally flatter, and more consistent over their discourse, while /f/ productions tended to be less regular, with regions of higher energy appearing and disappearing. The data window for this analysis was set to 512-points (10.6 ms).

In order to attempt to capture this apparent difference in productions, the mean spectral peak

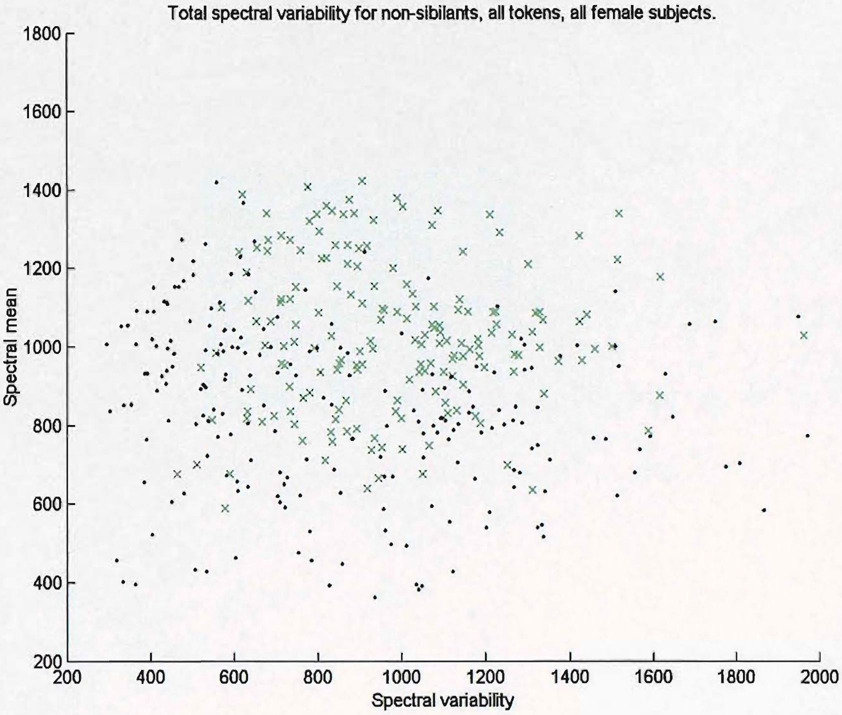


FIGURE 6.18: Scatter plot showing mean spectral energy against total spectral variance for all non-sibilant tokens, all female subjects. Green crosses are /f/ while black dots are /θ/.

frequency is calculated:

$$\upsilon_{[/F/,v,s,r]} = \overline{\left\{(f_\xi)_{[/F/,v,s,r]}\right\}} \tag{6.22}$$

$$= \frac{1}{N_w}\sum_{n=1}^{N_w}(f_\xi)_{[/F/,v,s,r,n]} \tag{6.23}$$

where the peak frequency of a given spectral window $(f_\xi)_{[/F/,v,s,r]}$ is defined

$$\hat{\Omega}(f_\xi)_{[/F/,v,s,r,n]} = \max_{f_k > 1\text{kHz}}\left\{\hat{\Omega}(f_k)_{[/F/,v,s,r,n]}\right\} \tag{6.24}$$

assuming $f_\xi$ is unique for all $f_k$. The variance of the spectral peak frequency through time was calculated

$$\varrho^2_{[/F/,v,s,r]} = \text{var}\left\{(f_\xi)_{[/F/,v,s,r]}\right\} \tag{6.25}$$

$$= \frac{1}{N_w - 1}\sum_{n=1}^{N_w}\left((f_\xi)_{[/F/,v,s,r,n]} - \overline{\left\{(f_\xi)_{[/F/,v,s,r]}\right\}}\right)^2. \tag{6.26}$$

Typical peak variability plots over time are shown in figures 6.19 and 6.20. These figures demonstrate nicely differences that can be observed in the non-sibilant multitaper spectrograms: /θ/ productions are commonly flat, and regular over time; this corresponds to a peak position that tends to jump around, since there is little difference in height of spectral maxima between spectra. However, /f/ productions often seem to have a dominant region of higher energy, and this

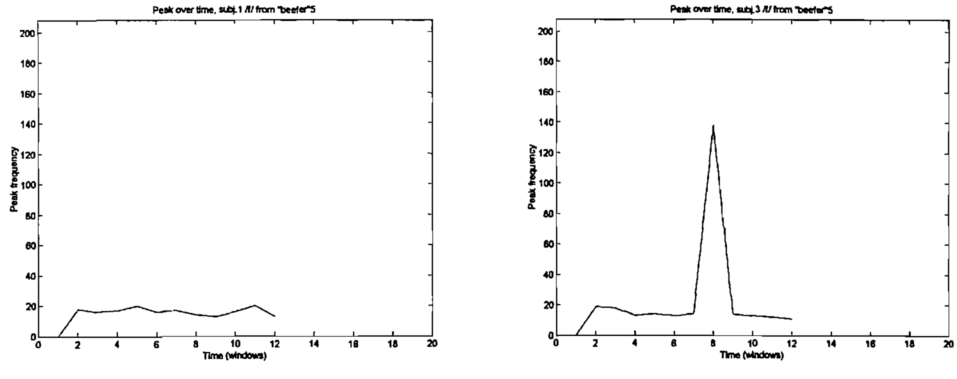


FIGURE 6.19: Examples of typical peak positions over time in [f] from /ifa/ subject M-01 (left) and M-03 (right).

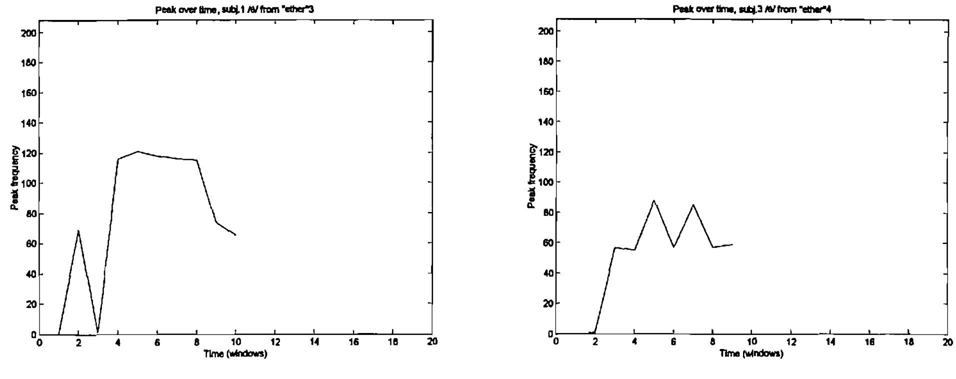


FIGURE 6.20: Examples of typical peak positions over time in [θ] from /iθa/ subject M-01 (left) and M-03 (right).

corresponds to a peak that tends to linger in only one or two regions.

A plot of spectral peak variability against mean spectral peak for male tokens of non-sibilants in all vowel contexts is shown in Figure 6.21. As expected, most of the /f/ tokens have very low peak frequency variability, and are clustered in the bottom-left. The /θ/ tokens are more evenly distributed over the variance scale. The number of tokens with low peak variability is clarified with the use of the histograms in Figure 6.22. The large majority of /f/ tokens have a very low (close to zero) peak frequency variability, while the majority of /θ/ tokens do not.

Of significant interest is that these results are almost the inverse as those discovered for the *spectral* variance found for the male non-sibilant tokens in §6.1.3: there, /f/ tokens were found to have the larger degree of variability, while using this different measure, they are found to have the smaller variability. This however, adds to the evidence that information pertaining to the place of production is inherent in the variability of the spectrum and its features over time.

Peak variability results for female non-sibilant tokens are shown in Figure 6.23. Once again, the female tokens do not follow the trend noticed in the male tokens strictly, (although there still are more /f/ tokens with near-zero spectral peak variance than for /θ/), but this is overshadowed by the large number of tokens with very large spectral peak variability for both non-sibilants, as highlighted by the histogram plots in Figure 6.24.

These measures of variability can of course be applied to the sibilants, whose structure is generally

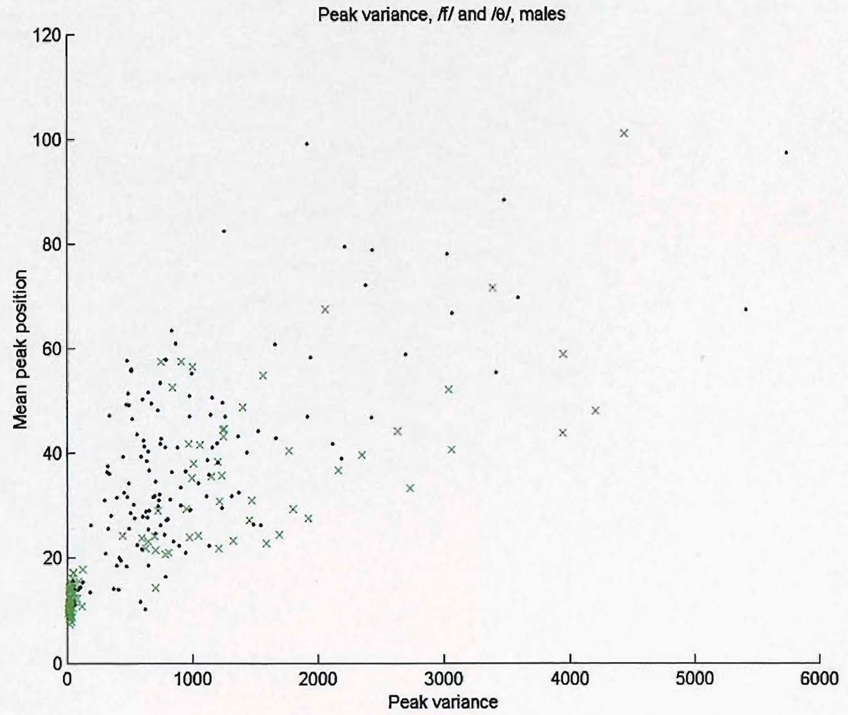


FIGURE 6.21: Measures of peak variability for all non-sibilants in all contexts by all male subjects. Green crosses are /f/, black dots are /θ/.

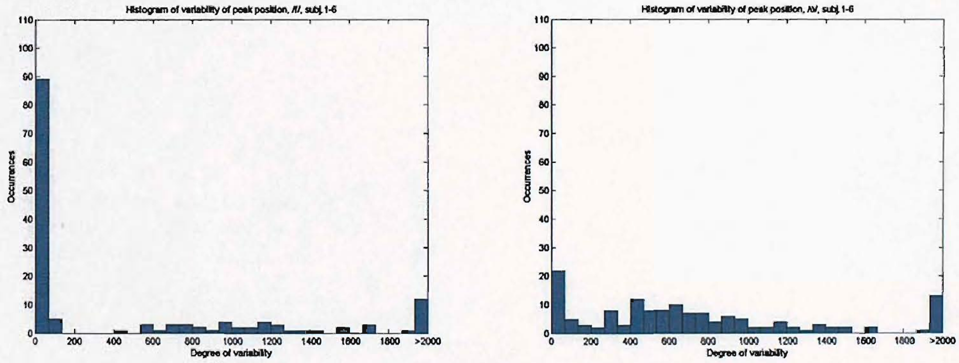


FIGURE 6.22: Histogram of variability of peak frequency, versus frequency, for /f/ (left) and /θ/ (right) in all contexts by all male subjects.

more predictable. Nevertheless, the results may be of interest. The peak variabilities for sibilant tokens is shown in Figure 6.25

The results for this analysis of spectral peak frequency variance, separated by speaker and by vowel context (again, for /iFi/ and /əFu/ contexts) are presented in §F.5. Again, the vowel context appears not to have a significant or consistent impact on the degree of variability.

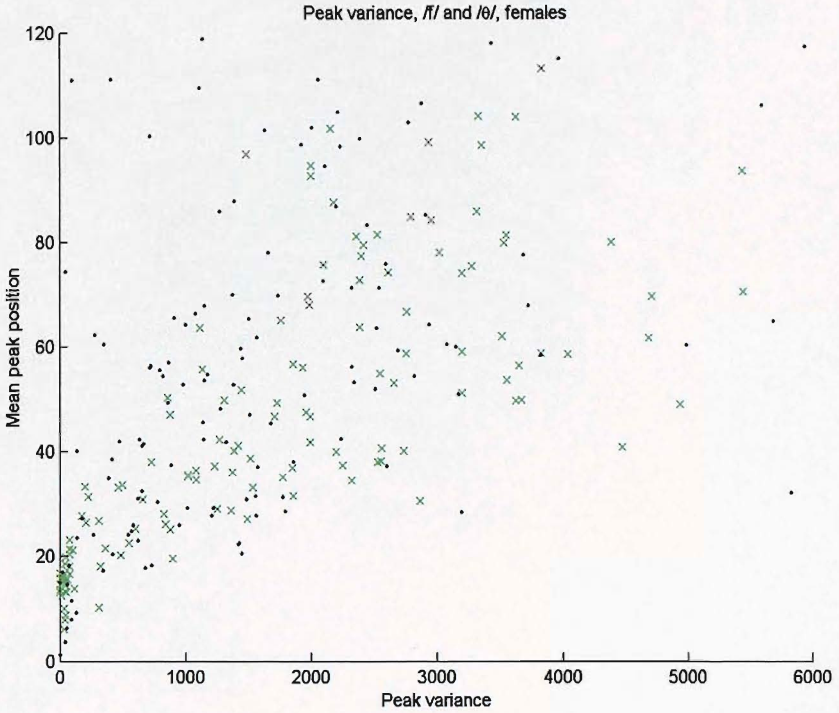


FIGURE 6.23: Measures of peak variability for all non-sibilants in all contexts by all female subjects. Green crosses are /f/, black dots are /θ/.

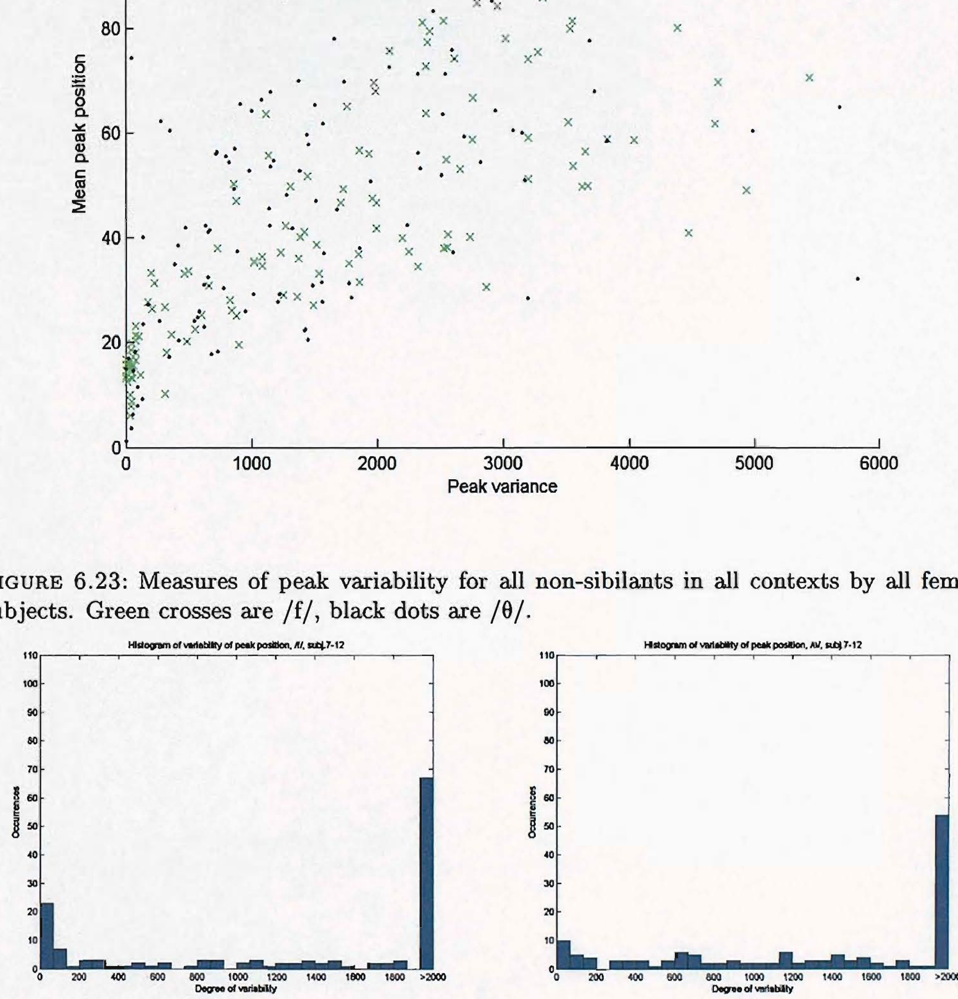


FIGURE 6.24: Histogram of variability of peak frequency, versus frequency, for /f/ (left) and /θ/ (right) in all contexts by all female subjects.

## 6.2 Spectral covariance

So far, our analysis has been restricted to estimating the variance as a function of frequency. It is also possible to calculate the covariance of the spectra at any two frequencies since this may reveal within-spectrum dependencies.

In order to continue the analysis of the voiceless fricatives in the vowel contexts considered so

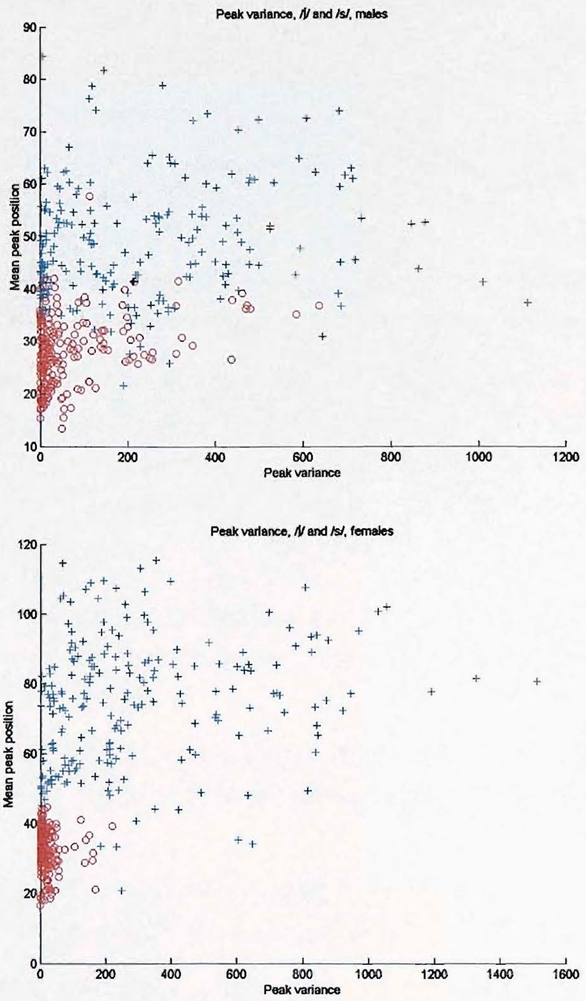


FIGURE 6.25: Measures of peak variability for all sibilants in all contexts by all male (top) and female (bottom) subjects. Blue crosses are /s/, red circles are /ʃ/.

far, spectral covariance matrices can be constructed:

$$\mathbf{c}_{[g,/\mathrm{F}/]} = \operatorname{cov}\left\{\hat{\Omega}(\omega_1),\hat{\Omega}(\omega_2)\right\}_{[g,/\mathrm{F}/]} \tag{6.27}$$

$$= \frac{1}{1295}\sum_{v=1}^{6}\sum_{s=1}^{6}\sum_{r=1}^{6}\sum_{n=1}^{6}\left(\left(\hat{\Omega}(\omega_1)\right)_{[/\mathrm{F}/,v,s,r]} - \overline{\left\{\hat{\Omega}(\omega_1)\right\}}_{[g,/\mathrm{F}/]}\right) \cdot \left(\left(\hat{\Omega}(\omega_2)\right)_{[/\mathrm{F}/,v,s,r]} - \overline{\left\{\hat{\Omega}(\omega_2)\right\}}_{[g,/\mathrm{F}/]}\right) \tag{6.28}$$

and the spectral correlation coefficient matrix is then defined

$$\mathbf{r}_{[g,/\mathrm{F}/]} = R\left\{\hat{\Omega}(\omega_1),\hat{\Omega}(\omega_2)\right\}_{[g,/\mathrm{F}/]} \tag{6.29}$$

$$= \frac{\operatorname{cov}\left\{\hat{\Omega}(\omega_1),\hat{\Omega}(\omega_2)\right\}_{[g,/\mathrm{F}/]}}{\sqrt{\operatorname{cov}\left\{\hat{\Omega}(\omega_1),\hat{\Omega}(\omega_1)\right\}_{[g,/\mathrm{F}/]}\operatorname{cov}\left\{\hat{\Omega}(\omega_2),\hat{\Omega}(\omega_2)\right\}_{[g,/\mathrm{F}/]}}}. \tag{6.30}$$

The window size for calculating the spectral covariance was set to 512-points (10.6 ms). The correlation of the spectrum to the total spectral energy can be calculated using this procedure. Appendix D shows the plots of correlation of total spectral intensity to spectral distribution for all male subjects producing all voiceless fricatives in all vowel contexts, both for combined speakers, and by speaker. Perhaps unsurprisingly, we find that the spectral distribution has a fairly linear relationship with total spectral intensity: as the intensity of the fricative increases, the power across the entire spectrum rises by a proportional amount, all the way up to 20 kHz. Occasionally, frequencies below about 2 kHz exhibit less of a dependence on the total intensity, most notably for the sibilants. These plots are otherwise unhelpful, and so we move on.

Appendix E show plots of the spectral correlation coefficient matrices for all voiceless fricative tokens produced by male subjects; again, starting with results for combined speakers, and then by individual speaker. (The calculation of the plots for individual speakers is similar to (6.28) and (6.30), except that the sum over subjects is removed). Firstly, consider how these matrices should be interpreted:

- Large values are represented by 'warmer' colours, red indicating the values approaching unity, while the 'cooler' colours represent the lower value, dark blue being the closest to zero.
- The correlation coefficients are the normalised covariance values. The matrix gives a measure of the probability that two variables are correlated. Thus, the main diagonal is always equal to unity.
- Values near the main diagonal tend to have high values, since the energy distribution in most areas of the fricative is in the form of narrow bands of energy.
- Very small squares on the diagonal that rapidly change to low-values indicate narrow bands of energy that are independent of nearby values, but which generally moves as a single small block.
- A large-valued square centred on the diagonal therefore most likely indicates that the energy in the range of frequencies in this band generally moves as a single block.

While these spectral correlations are very interesting, they unfortunately exhibit little distinguishing information. General trends are difficult to spot, although when considered individually, they reveal a certain amount about the individual speaker's productions of a particular fricative.

Generally speaking, the sibilants appear to have much lower degree of spectral correlation: the energy in one part of the spectrum will not generally be indicative of the amount of energy in another part. Conversely, the fricatives appear to have a slightly higher degree of spectral correlation, particularly in the form of a few broad bands of energy that seem to 'adhere' together.

Female data for these spectral correlations are not presented, since they generally provide very similar information. Plots of within-token spectral correlations were also calculated, using smaller (64-point) data windows, and hence at a coarser frequency resolution. However, these plots did not yield obvious place information, and so have not been included in the appendices. A more thorough analysis of such plots may reveal information of interest, although this was not undertaken in this work.

## 6.3 Discussion

We have shown that the use of multitaper estimates in the spectral analysis of fricatives is beneficial. The reduced variance of the estimate allows the estimation of spectral variation across productions, providing us with a very clear picture of distinguishing characteristics within the sibilants.

Analysis of spectral variance has been accomplished through the implementation of multitaper spectral estimation. Such an analysis could not be performed using basic modified periodogram spectral estimates, since the variance of the estimate would have swamped the underlying spectral variances. Furthermore, time and ensemble averaging techniques would also make such an analysis almost impossible, due to the very nature of the underlying assumptions about the process. This analysis of variance then, potentially provides us with new information pertaining to fricative production.

Using these techniques, it is straightforward to locate regions of the spectrum that are highly variable, and those that are more stable, across speakers, vowel contexts, or even individual tokens. It was quickly discovered that the region around 2.5 kHz for male tokens, and the 3-kHz region for female tokens, is both one of low variance, and of a distinguishing feature of the sibilants. Coupled with the total spectral energy with which this region of the spectrum is loosely correlated, strong evidence has been found that these regions may be a distinguishing feature of the two sibilants, and figures 6.5 and 6.8 describe these findings most effectively.

It is *not* asserted that our analysis serves to prove the effectiveness of these measures as general classification metrics, since no suitable statistical tests have been performed with which such a statement could be qualified. Indeed, the number of subjects, and the size of corpus — while extensive for the purposes of this study — are probably not sufficient to claim accurate knowledge of some larger data set. Nevertheless, the usefulness of the examined techniques cannot be refuted. The results obtained are both intuitive, and also fit current general theories on fricative production.

Spectral variance analysis clearly demonstrated the limitation of attempting to capture differences in the non-sibilants based on spectral shape. The expected variability of the spectral shape over any arbitrary mid-fricative segment is often large, and spectral features that consistently distinguish the non-sibilants are not apparent. Nevertheless, significant trends have been observed: most /f/ tokens have spectral peak position near 2 kHz. When not located at 2 kHz, the peak tends to occur near 4 kHz, or 7 kHz for males, or 8 kHz for females. Generally, /θ/ tokens also often have their main spectral peak near to 2 kHz, but more often than for /f/ tokens, it is elsewhere, commonly around 6 kHz for males, but otherwise more evenly distributed.

Evidence has also been found to suggest that the variability of non-sibilant tokens over time may show different characteristics for /θ/ and /f/. For non-sibilant tokens produced by males, it appeared that the total amount of spectral variability over the total length of a given token was often lower for /θ/ tokens. Tracking the spectral peak frequency through non-sibilant tokens also provides evidence that differences between the non-sibilants exists in the form of spectral variations over time. Categorising these results by vowel context reveals little additional information.

Plots of spectral covariance also provide a new representation to view characteristics of fricative tokens. While distinguishing trends are hard to pinpoint, the spectral covariance plots seem to provide information about interdependencies that exist within the spectrum during fricative production.

Evidence now exists that information pertaining to place can be discovered by using good spectral estimation techniques that minimise estimate variance, in order to measure variations in the underlying process spectrum. Moreover, these measurements have a highly intuitive appeal in comparison to, say, spectral moments. Comparing these results to the best spectral moment results in Chapter 5 demonstrates the importance of using good spectral estimation techniques prior to establishing characterisation metrics. It is hoped that these new techniques can be applied to fricative tokens produced by disordered speakers.

# Chapter 7

# Preliminary measurements of cochlear implant users

One of the main motivating factors for improving fricative production analysis is that of better describing the fricative production of speakers with hearing that has been changed. In this way, it is hoped that subtle changes to speech quality that are brought about by changes in hearing can be measured more accurately.

In this chapter, examples of fricative productions of two male and two female cochlear implant subjects are examined. In this way, the advantages and disadvantages of the various analysis techniques can be compared.

We begin by analysing the male subject data, using classical analysis techniques such as spectrograms and spectral moments with the most reliable parameters as discussed in Chapters 4 and 5. We continue to compare these analysis methods with some of the new methods developed in Chapter 6.

The female subjects provide some more diverse productions, and the same methods of analysis are undertaken. Data for subject FCI-15 is taken from 1 year post, and 2 years post implantation, since a comparison over this interval reveals a significant change in production of /s/.

## 7.1 Male productions

Multitaper spectrograms for productions of /ɑsə/ from 'parcel' are given for both male cochlear implant subjects in figures 7.1. From these plots, the productions appear similar to those for normal hearing subjects, as can be seen in Appendix B: lower energy at low frequencies around 3 kHz, but quite high energy, at least in the 4–16-kHz range. While these two productions are quite dissimilar from each other, they are typical within the range of variation we have come to expect among the sibilants, at least at all frequencies other than around 3 kHz; and at this frequency we note low energy for both.

Figure 7.2 shows two non-sibilant tokens for these two speakers. Again, the spectrograms appear

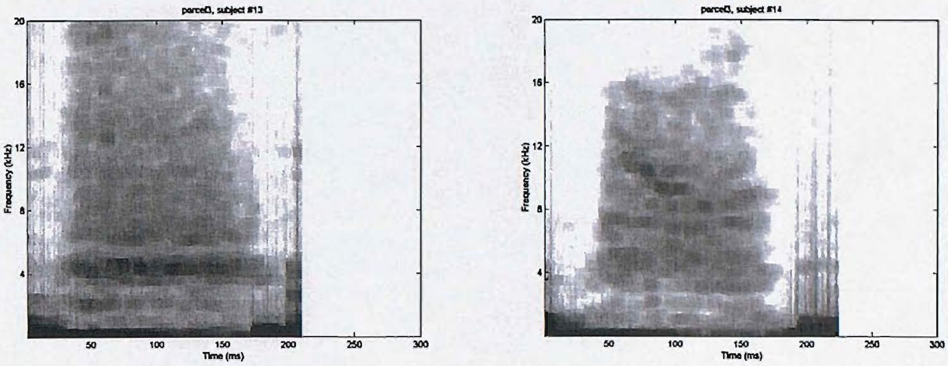


FIGURE 7.1: Multitaper spectrograms of /ɑsə/ productions from 'parcel', subject MCI-13 (left) and MCI-14 (right).

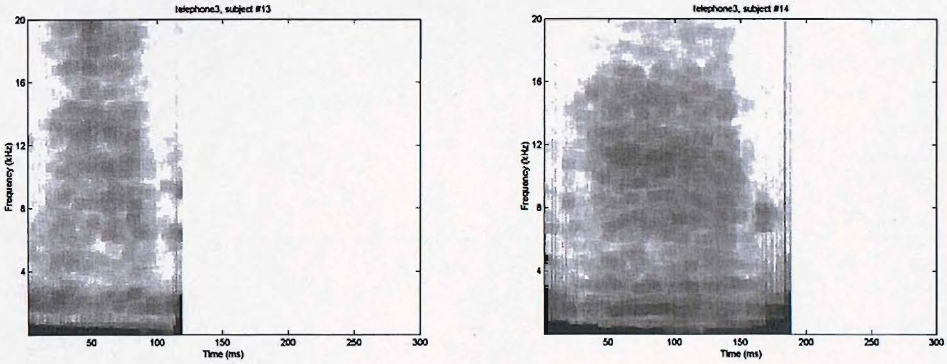


FIGURE 7.2: Multitaper spectrograms of /əfəʊ/ productions from 'telephone', subject MCI-13 (left) and MCI-14 (right).

to be similar to the productions from 'normal' hearing subjects shown in Appendix B. Figure 7.3 shows the calculated first and second spectral moments using the methods described in Chapter 5. An interval 64 ms long located mid-fricative was used to calculate each mean spectrum from which the spectral moments could be calculated. These spectral moments can be compared to those in Figure 5.9. Principal component analysis (PCA) has been used to construct ellipses that enclose 85.35% of the data points for 'normal' male productions. The /s/ production of MCI-13 has rather a low centroid value compared to the productions of normal hearing male subjects, suggesting this production is more /ʃ/-like. The second moment for this production is within satisfactory limits however. The /s/ production of MCI-14 is within the ranges suggested by the normal hearing male subjects, although the second moment is quite high. The /f/ tokens for both male cochlear implant subjects appear to have normal centroid values, but rather high second moment values.

Plots of total spectral energy against peak location, calculated from 64-ms mid-fricative data segments for these tokens are given in Figure 7.4. These plots can be compared to those of the normal-hearing subjects in Figure 6.10. The plots suggest spectral peak position values similar to those found for normal-hearing speakers. The /f/ token of subject MCI-13 is also at 2 kHz, suggesting good similarity to 'typical' productions. The /f/ production for MCI-14 has a high-frequency spectral peak, but is still within 'normal' limits.

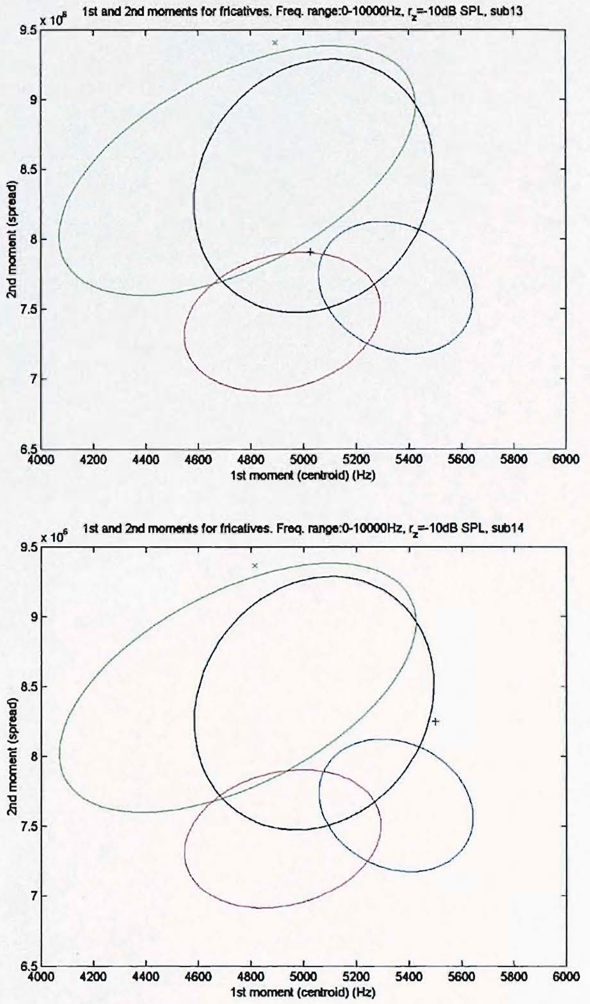


FIGURE 7.3: 1st and 2nd moments of /s/ (blue plus) and /f/ (green cross), subjects MCI-13 (top) and MCI-14 (bottom). PCA has been used to construct ellipses that enclose 85% of the 'normal' production points: /ʃ/ blue, /s/ red, /f/ green, /θ/ black (see Figure 5.9).

The energies in the 2.5-kHz band for each of the /s/ tokens are within satisfactory limits. The /s/ token of MCI-13 lies near the top-right-hand corner of Figure 6.5, suggesting a typical 'louder' production. The /s/ token of MCI-14 is quieter, and has correspondingly lower energy in the 2.5-kHz band, ending up nearer the bottom-left-hand corner of the distribution in Figure 6.5.

The mean spectra, and the spectral variance calculated over each fricative token, as described by equations (6.17) and (6.19) in §6.1.6, were calculated for fricative tokens by male cochlear implant subjects, and are shown in figures 7.5 and 7.6. These plots again suggest that the spectral variability over the duration of the fricatives of these subjects, is similar to those seen in the productions of normal-hearing subjects (which can be found in §F.3.1), except perhaps the /s/ of MCI-14. This /s/ token shows level of spectral variability somewhat higher than the /s/ tokens produced by the normal hearing male subjects. In fact, similar spectral variance distributions can be found in some of the /ʃ/ tokens of the normal hearing male subjects (e.g. M-05 in Figure F.28). This may suggest that this /s/ token has some production characteristics that are closer to /ʃ/, and although such an observation is largely speculative, it remains that

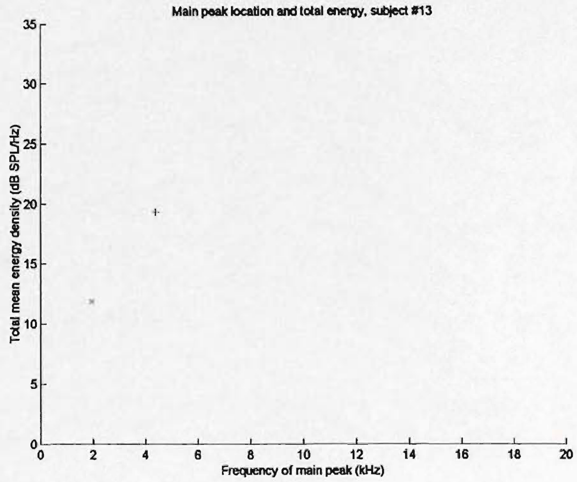
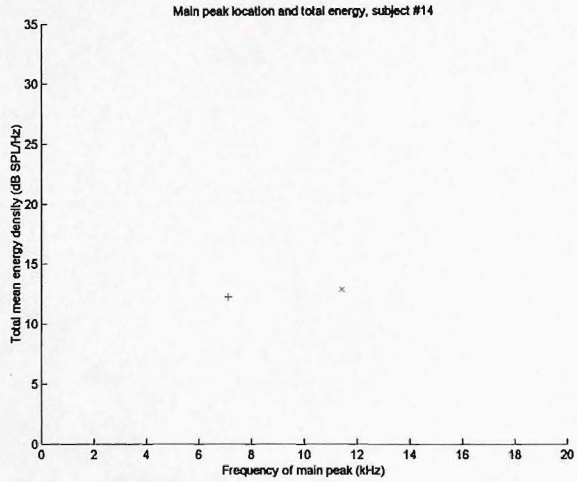



FIGURE 7.4: Total spectral energy and peak frequency of /s/ (blue plus) and /f/ (green cross) tokens, subjects MCI-13 (top) and MCI-14 (bottom).

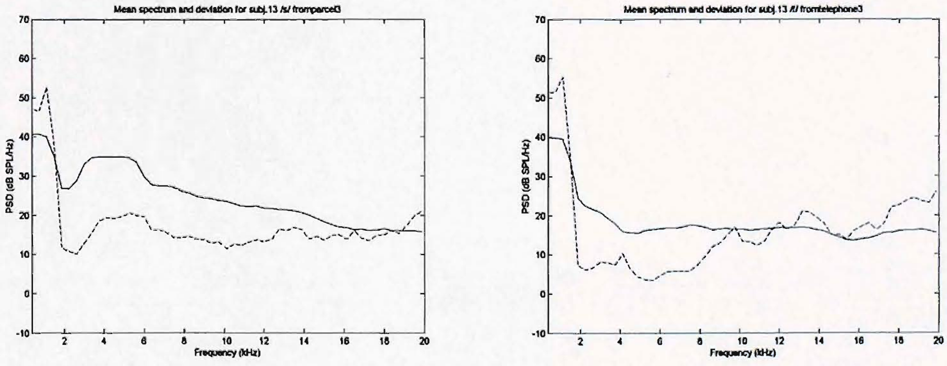


FIGURE 7.5: Fricative mean spectrum (solid line), and spectral variance (dashed) over /ɑsə/ productions from 'parcel' (left), and /əfəʊ/ productions from 'telephone' (right), MCI-13.

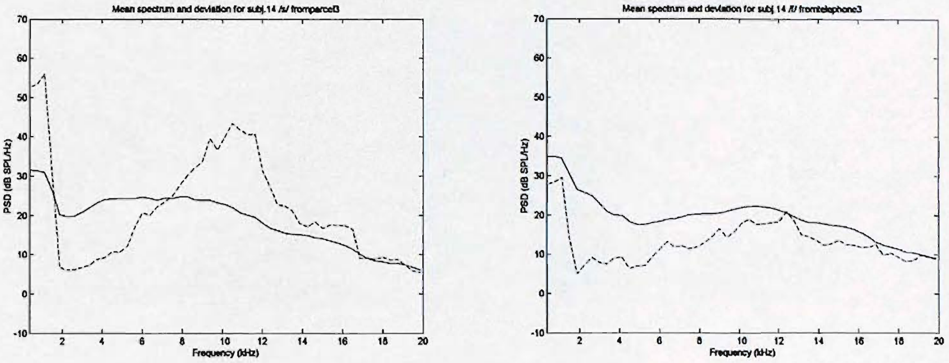


FIGURE 7.6: Fricative mean spectrum (solid line), and spectral variance (dashed) over /ɑsə/ productions from 'parcel' (left), and /əfəʊ/ productions from 'telephone' (right), MCI-14.

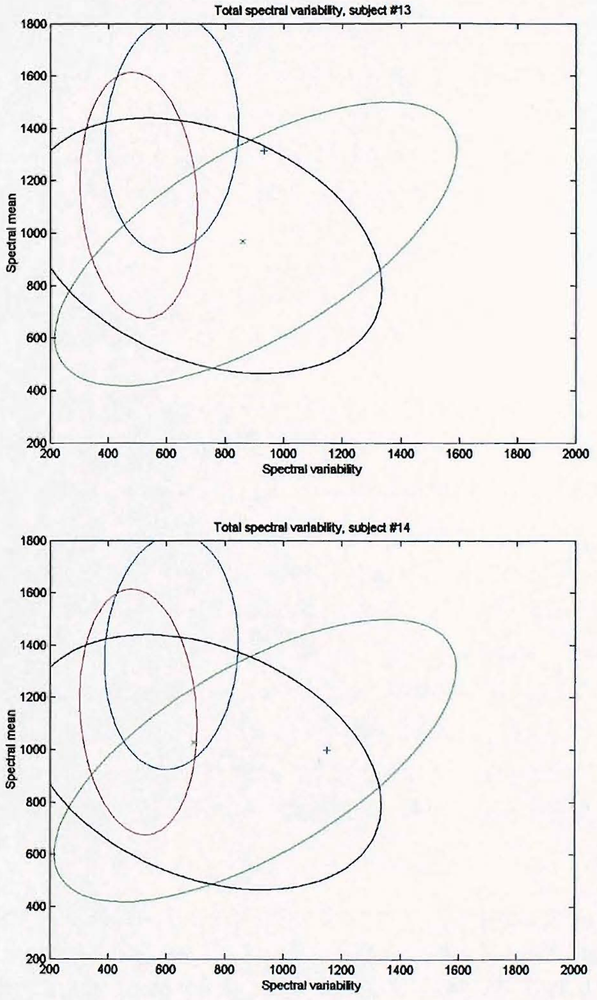


FIGURE 7.7: Spectral mean and total spectral variance over time, subjects MCI-13 (top) and MCI-14 (bottom). PCA has been used to construct ellipses that enclose 85% of the 'normal' production points: /ʃ/ blue, /s/ red, /f/ green, /θ/ black. Legend: /s/ blue plus, /f/ green cross.

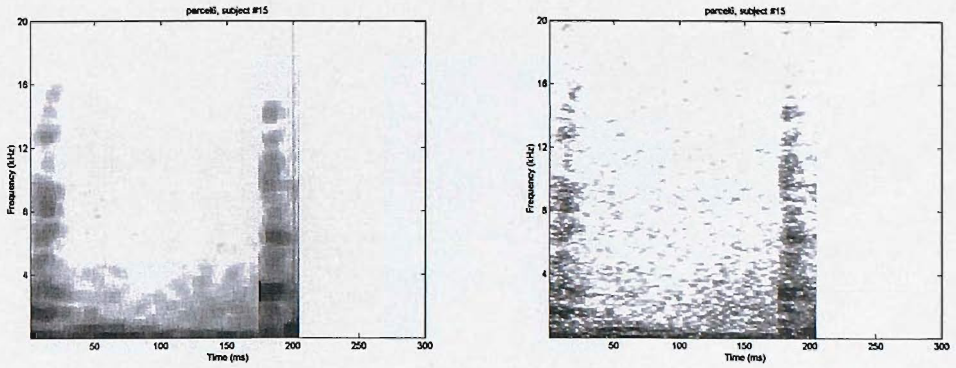


FIGURE 7.8: /ɑsə/ productions from 'parcel', subject FCI-15, one year post implant. Multitaper spectrogram (left) and traditional modified periodogram spectrogram (right).

the /s/ production most certainly appears to have a greater degree of variability than might normally be expected. Scatter plots of mean spectrum, and total spectral variability are given in Figure 7.7, and these can be compared to the non-sibilant total spectral variance plot in Figure 6.17. For subject MCI-13, the /f/ token lies somewhere near the centre of most /f/ tokens for normal hearing male subjects, and is well outside the region more commonly occupied by /θ/. The /f/ token of MCI-14 also lies in the centre of typical /f/ tokens by the normal hearing males, although it is within the region of significant overlap with /θ/ tokens.

## 7.2 Female productions

We now consider a production of /ɑsə/ by subject FCI-15, approximately one year after implant insertion. A multitaper spectrogram and traditional spectrogram are shown in Figure 7.8. It is straightforward to observe that the production by subject FCI-15 has failed to produce any significant energy above around 500 Hz. The multitaper spectrogram correctly shows no energy in the upper frequency region. However, careful observation of the traditional spectrogram reveals small 'patches' of energy up to around 16 kHz; this is clearly misleading, since total closure has occurred for this production. In fact, these patches of energy are due to the low level background noise, which of course can also be considered a stochastic process. A modified periodogram spectral estimate of this background noise will inevitably result in erroneous 'spikes' of energy that are simply a result of the large error variance. Since the multitaper periodogram is better suited to representing this noise, it provides a slightly clearer picture of the production in this case.

A production of /ɑsə/ by subject FCI-15 one year later, at two years post-implantation, is shown in Figure 7.9. It appears that the production of this token has improved over this year-long interval. Frication noise is significant up to approximately 18 kHz. Closer analysis also reveals that the main spectral peak is possibly at a rather low frequency for a typical /s/. Also, there is a 'pause' mid fricative, at which point frication noise ceases momentarily.

Figure 7.10 shows the multitaper spectrograms for /f/ productions of the female cochlear implant subjects. While the production of FCI-16 appears similar to those produced by normal hearing

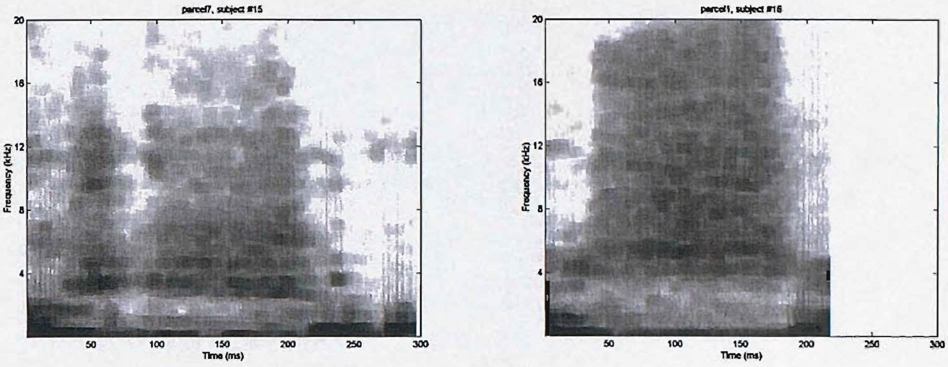


FIGURE 7.9: Multitaper spectrogram of /ɑsə/ production from 'parcel', subject FCI-15 two years post implant (left), and subject FCI-16 one year post implant (right).

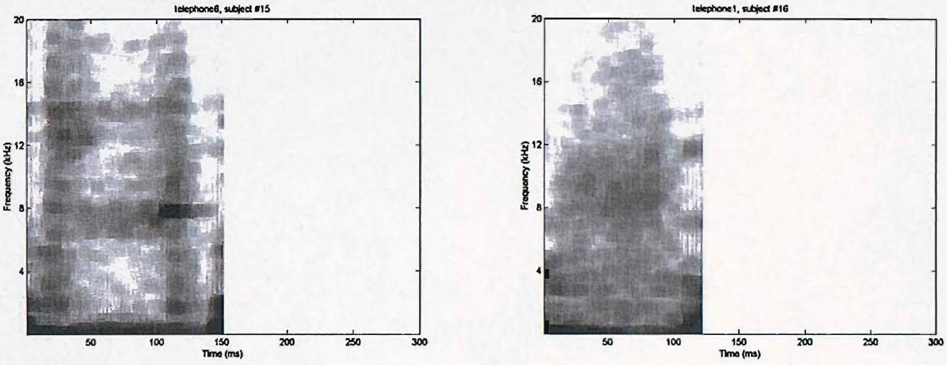


FIGURE 7.10: Multitaper spectrograms of /əfəʊ/ productions from 'telephone', subject FCI-15 (left) and FCI-16 (right).

subjects (see §B.2), the token of FCI-15 again has a strange temporary attenuation of frication noise mid-fricative. On listening to this production, the fricative is apparently whistled.

We now take a moment to consider the results obtained when various analysis methods for fricatives are used on these types of 'abnormal', highly nonstationary productions on various analysis methods.

Figure 7.11 shows the first and second spectral moments calculated for each of the female fricative tokens. Consider first the /s/ of subject FCI-15, the centroid of the spectrum is around 5 kHz, while the second moment is around $7.3 \times 10^6\,\mathrm{Hz}^2$. Comparing these results to those of normal hearing female subjects in Figure 5.9 it can be seen that this production lies within the range normally associated with female /ʃ/ tokens, and indeed, this corresponds to the subject production notes for FCI-15. The /s/ token of FCI-16 lies well within the range corresponding to productions by the female normal-hearing subjects. The /f/ tokens also lie within 'normal' limits.

Scatter plots of total spectral energy, and peak frequency for the female cochlear implant subjects are shown in Figure 7.12. These plots can be compared to those in the plot for normal-hearing females in Figure 6.11. The spectral peak of the /s/ production by FCI-15 occurs within the /ʃ/ region defined by the normal-hearing female subjects, and this agrees with the /ʃ/-like spectral moment values for this token. To serve as a comparison, an example /ʃ/ token selected manually

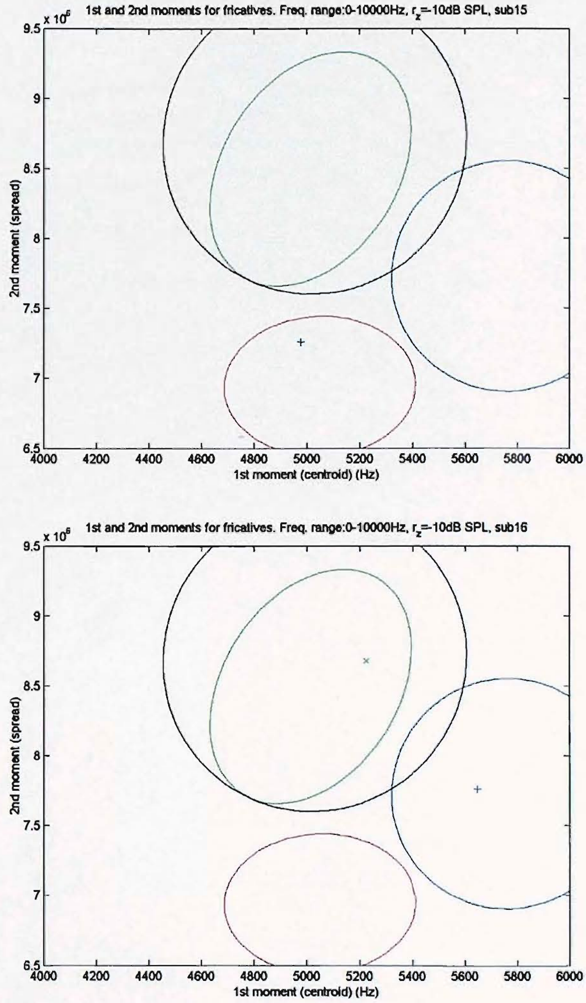


FIGURE 7.11: 1st and 2nd moments of /s/ (blue plus) and /f/ (green cross), subjects FCI-15 (top) and FCI-16 (bottom). PCA has been used to construct ellipses that enclose 85% of the 'normal' production points: /ʃ/ blue, /s/ red, /f/ green, /θ/ black (see Figure 5.9).

from a production of the word 'shoe' by FCI-15 is also shown in the upper plot of Figure 6.11; this /ʃ/ token lies within the region heavily populated by [ʃ] productions of normal hearing subjects, and highlights the similarity in spectral peak position between these two sibilant productions of subject FCI-15. Figure 7.13 demonstrates again the similarity between these two productions, using the energy at 3 kHz against total spectral energy measure that was able to completely separate the normal hearing female sibilant productions (c.f. Figure 6.8).

From the lower plot in Figure 7.12, it can be seen that the 6-kHz peak frequency of the /s/ token of FCI-16 however, lies well within the values defined for typical normal-hearing female [s] tokens. In both plots for the female cochlear implant subjects, the /f/ tokens have spectral peaks near 8 kHz. This also corresponds to one of the peak positions often observed for normal hearing female /f/ tokens when it does not occur at 2 kHz.

Plots of mean spectrum, and spectral variance over the duration of the female cochlear implant fricative tokens — as calculated in equations (6.17) and (6.19) — are given in figures 7.14 and 7.15. The /s/ token plot in Figure 7.14 excludes the obvious pause mid-/s/ production for

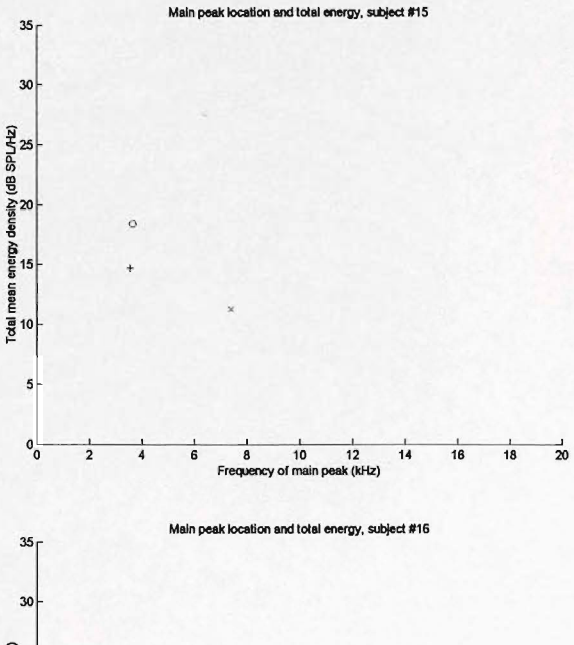


FIGURE 7.12: Total spectral energy and peak frequency of /s/ (blue plus) and /f/ (green cross) tokens, subjects FCI-15 (top) and FCI-16 (bottom). Top plot includes results for example /ʃ/ token (red circle) from 'shoe' for FCI-15.

FCI-15 (including it simply raised the spectral variance, as expected). We see that indeed, the mean spectrum resembles that of a typical /ʃ/ spectrum for the normal hearing female subject tokens. The spectral variance shown for the /f/ token of FCI-15 also shows a very high level in the 6–10-kHz region, compared to those typically seen in productions by normal-hearing females (see §F.3.2); although occasionally the spectral variance can be high among the normal hearing female /f/ tokens, it is rarely as large as for this 'whistled' production.

The mean spectrum of the /s/ token of FCI-16 appears more similar to those of normal female /s/ spectra, but on this occasion, a much higher degree of spectral variance has occurred than typically found amongst /s/ tokens by the normal hearing female subjects.

Scatter plots of total spectral variance and mean spectrum are shown for the female cochlear implant tokens in Figure 7.16. The value for non-sibilants can be compared to those calculated for the normal hearing female tokens in Figure 6.18. While the values for the /f/ token of FCI-15 are within comparable values, the production of FCI-16 appears to be within the range of values normally only occupied by /θ/ tokens of the normal hearing female subjects.

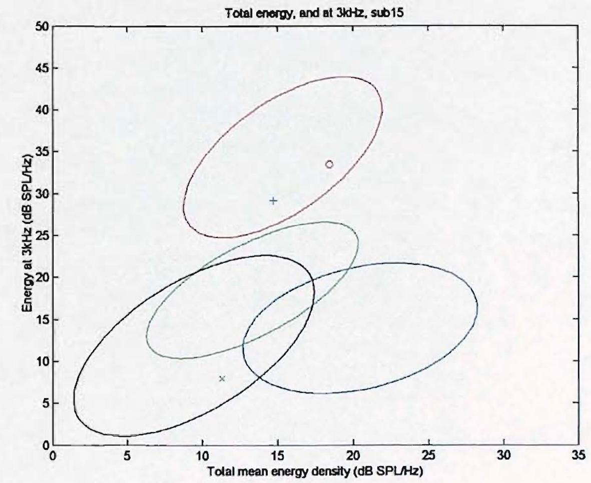


FIGURE 7.13: Energy at 3 kHz against total spectral energy for productions of /s/ (blue plus) and /ʃ/ (red circle) from subject FCI-15. PCA has been used to construct ellipses that enclose 85% of the 'normal' production points: /ʃ/ blue, /s/ red, /f/ green, /θ/ black.

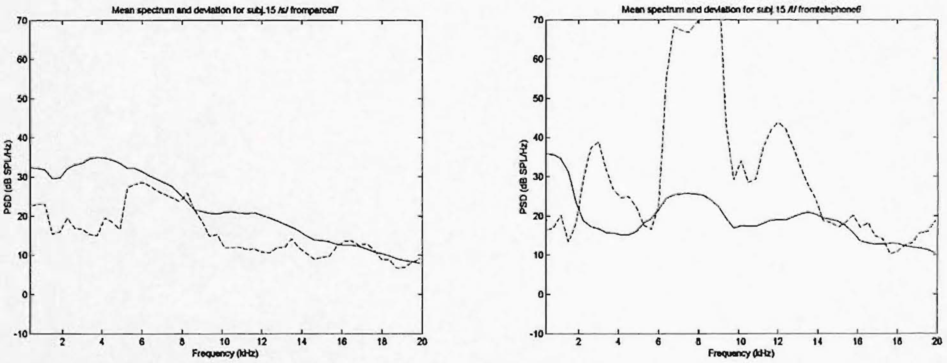


FIGURE 7.14: Fricative mean spectrum (solid line), and spectral variance (dashed) over /ɑsə/ productions from 'parcel', and /əfəʊ/ productions from 'telephone', FCI-15.

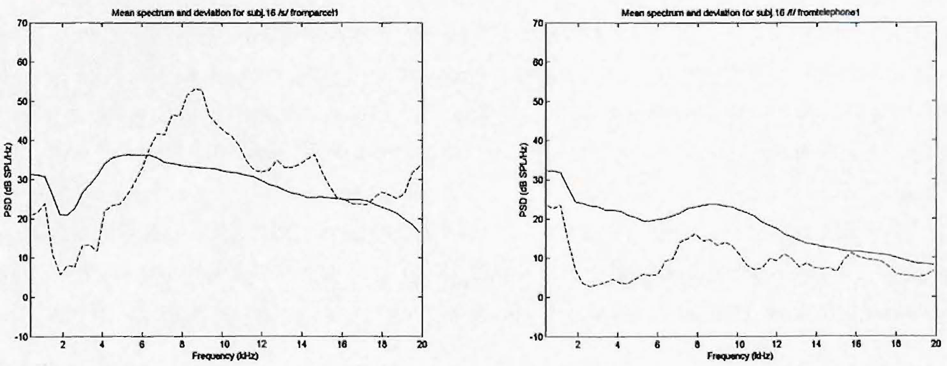


FIGURE 7.15: Fricative mean spectrum (solid line), and spectral variance (dashed) over /ɑsə/ productions from 'parcel', and /əfəʊ/ productions from 'telephone', FCI-16.

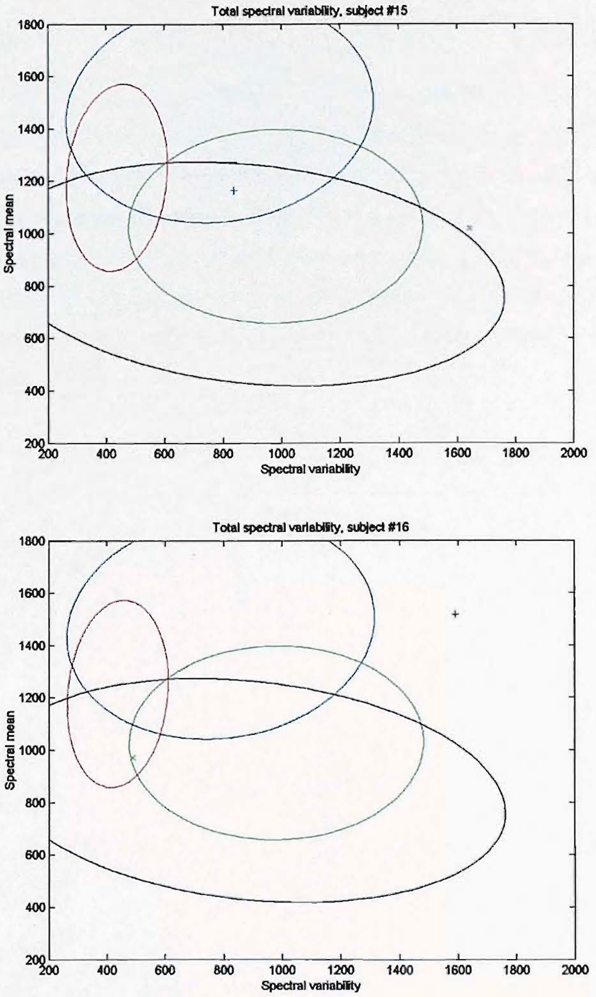


FIGURE 7.16: Total spectral variance over time, subjects FCI-15 (top) and FCI-16 (bottom). Legend: /s/ (blue plus), /f/ (green cross). PCA has been used to construct ellipses that enclose 85% of the 'normal' production points: /ʃ/ blue, /s/ red, /f/ green, /θ/ black

## 7.3 Summary

Two of the most established methods of fricative analysis are the spectrogram, and spectral moments. A number of improvements have been made to each method, and these have been shown to generally help produce slightly clearer results, which is especially important when considering disordered speech. Erroneous patches of energy in spectrograms generated using modified periodograms are eliminated with the aid of multitaper estimates. Our knowledge of suitable frequency and amplitude scales for spectral moment calculation in order to maximise the distance between sibilant clusters, and minimise correlation amongst the moments, mean that calculations of disordered productions are given the best chance of accurate description.

A number of new methods have also been able to be developed using improvements in spectral estimation from multitaper analysis. Measures of the spectral variations occurring over the duration of fricative tokens are now possible, and provide interesting and intuitive information about individual fricative productions. Moreover, these new analysis techniques seem to provide a greater diversity of information, and in several cases, have yielded evidence of 'abnormalities' in productions that appeared to have 'normal' spectral moment values.

At the very least, these new measures seem to provide invaluable supplementary information about productions. Moreover, they may provide useful information on their own, and since they deal more elegantly with nonstationary elements within fricatives than spectral moments, they may be favoured where disordered productions are leading to temporal features. Additionally, the measures are more easily interpretable: spectral peak location, energy density at certain frequencies, and measures of the degree of change over the course of a production can be loosely related to the underlying acoustical mechanisms, and with further work, these relationships should begin to become clearer.

# Chapter 8

# Conclusion

## 8.1 Summary

Fricative analysis presents a significant challenge. Too little is known about the turbulent noise sources that are generated within the tract during fricative production. The interactions of multiple noise sources are largely unknown, and usually effectively impossible to calculate. Mathematical models of fricative production inevitably over-simplify the processes within the tract. While much has been learnt from such models, their usefulness when applied to fricative production is limited.

Nonparametric measurements of the output of the system are likely to provide useful information pertaining to various characteristics of production. Since a certain amount is known of the characteristics of the acoustical signal generated during fricative production, methods of analysis should incorporate these characteristics where possible. For example, it is known that the peaks in the spectrum correspond to resonances within the tract, and so studying the behaviour of these is more likely to lead to better understanding of fricative production that other measurements that are less grounded on the physics of production.

### 8.1.1 Classic spectral estimation techniques

The turbulence noise generated during fricative production should be treated as a stochastic process. Yet often in the fricative analysis literature this is overlooked, despite well founded texts on the issue (e.g. Bendat and Piersol 1986). While windowing of time-series data (to reduce the spectral bias) is usually performed, it is not usual to attempt to reduce the variance error of the estimate, which is generally large. For the purposes of studying spectral peak positions and so on, a modified periodogram estimate is unsatisfactory. Nevertheless, this estimate is commonly used in the fricative analysis literature.

In order to reduce the variance error of the estimate, several classic averaging methods exist. Time averaging can be used when the process can be considered stationary, and ensemble-averaging can be used where a process is ergodic. However, there is little to suggest that fricative production can be considered either stationary or ergodic. An alternative method of reducing the variance when these assumptions do not hold is frequency smoothing. However, this relies upon the underlying spectrum being smooth, and in order to generate an estimate that significantly reduces the variance error, it also reintroduces significant local bias.

### 8.1.2 Data acquisition

In order to gain a first estimate of typical variations across productions, analysis of some 'normal' speech was necessary. A large corpus of real words containing each of the eight English fricatives in /$V_1FV_2$/ contexts has been devised. Six vowel contexts are incorporated, and words were repeated six times, each time in a slightly different word order. This corpus was read by six male, and six female normal hearing subjects of Southern English accent background. This resulted in a data set of 1,728 voiced, and 1,728 unvoiced fricative tokens. However, only the voiceless fricative tokens are considered here.

To supplement these data, the speech of two male and two female post-lingually deafened cochlear implant subjects was recorded. A real word corpus was used, so that a small set of fricative tokens from these subjects could be used to compare possibly disordered speech results to those of the normal hearing subjects.

### 8.1.3 Multitaper analysis

Multitaper analysis provides an alternative method of obtaining a spectral estimate with minimised error. The quantity of data incorporated into the estimate is maximised using the prolate spheroidal functions, or Slepian sequences, as data tapers on a short interval of time-series data. The local bias is minimised at the same time due to the specific properties of the Slepian sequences.

Most importantly, multitaper analysis does not rely upon assumptions of stationarity, or ergodicity. It therefore outperforms the alternative averaging methods, where consistent spectral estimation over short time intervals of non-stationary non-ergodic processes is required. These properties make multitaper analysis an excellent candidate for fricative spectral analysis.

Multitaper analysis enables changes in the spectrum over time to be observed more easily. Comparisons across productions can also be made. Finally, multitaper spectrograms can be generated that do not contain the 'speckle' usually found in spectrograms of fricatives, and in some cases, multitaper spectra are more straightforward to read.

### 8.1.4 Best results for spectral moments

Spectral moments provide a broad measure of the overall energy distribution within a spectrum. The Gram-Charlier expansion has been used to demonstrate the elements of spectra that are most influential to the calculation of spectral moments. They are not significantly influenced by the movements of narrow spectral peaks or troughs. Rather, they are more sensitive to significant changes in energy distribution.

A number of choices concerning the implementation of spectral moments have to be made, most significantly frequency range selection, magnitude scale selection, and zero reference. These choices can significantly affect the effectiveness and sensitivity of the spectral moments. Historically, various approaches have been implemented, but often with little reasoning given. It has been shown that in fact, these choices can make significant differences to the outcomes of spectral moment calculations. If the zero reference is set too low, the spectral moments become insensitive to changes in the energy distribution of the spectrum. If it is set too high, there is an increased risk of the spectrum dropping below it, generating spurious moment results.

The spectral moment methodology was adapted to use multitaper spectral estimates, rather than modified periodogram estimates. It seems probable that higher-order spectral moments are stabilised when multitaper spectral estimates are used, due to a decrease in variation in the tails of the distributions. The Gram-Charlier expansion allowed us to view which features of a spectral distribution most influence the spectral moments calculated for that distribution. Various frequency ranges were tried, and 0–10 kHz seems to produce among the best results. Logarithmic (decibel) magnitude scales should be used since the spectra have more Gaussian-like qualities. This highlights a further important parameter: the zero reference, above which the spectral distribution can be normalised. This normalisation procedure is required in order to calculate the spectral moments. The zero reference must not be set so high that some spectra drop below it. However, the sensitivity of the spectral moments to changes in spectral shape is reduced as the zero reference is lowered. A zero reference of −10dB SPL/Hz with a frequency range of 0–10 kHz is found to be a good balance that can deal with the significant degree of variation that exists across tokens; but these settings still result in a high degree of correlation between the even order moments, and between the odd order moments. It is found that the 1st and 2nd spectral moments provide the greatest amount of information, and have been shown to be capable of separating the voiceless sibilants well, although separation of the non-sibilants could not be achieved.

Plotting the spectral moments through time during fricative productions revealed evidence that the statistical properties of fricatives are often nonstationary. The non-sibilants were subject to the highest levels of spectral moment nonstationarity, although no distinguishing characteristics in these variations over time could be established. Nevertheless, this evidence of nonstationarity has serious repercussions concerning the use of time-averaging methods.

### 8.1.5 Analysis of spectral variance

Attempting to track the spectral changes that occur across productions, or through time within a fricative token, was previously difficult to perform accurately due to the large variance error of typical spectral estimates available. If attempts were made to reduce this variance, some assumption of stationarity or ergodicity would typically have to be made, and this still limits the accuracy of such analysis.

The use of multitaper spectral estimates therefore provides us with a new means of gathering information pertaining to fricative production. The variability across, and within tokens can be explored due to the much reduced variance of the estimate.

Multitaper spectral estimates were calculated from mid-fricative data. Spectra from six separate

10.6-ms windows within each mid-fricative token from every vowel context of every speaker for all speakers of a given gender were generated. The results were used to calculate mean spectral shapes, and spectral variances across all productions. The same analysis was performed to estimate within-speaker spectral mean and variance, and within-vowel-context and speaker spectral mean and variance. Finally, spectra were calculated from adjacent 2.7-ms data-windows across the whole duration of each fricative token. The spectral mean and variance over the duration of each fricative could then be estimated.

These analyses of production variabilities quickly reveal regions of the spectrum that are highly variable, and those that are more stable, across speakers, vowel contexts, or individual tokens. For example, it was discovered that the level of energy in the 2.5-kHz region in male sibilant productions was subject to a notably low level of variance. Additionally, this region coincides with the main spectral peak for /ʃ/ tokens, but a prominent trough for /s/ productions. The findings were similar for females, although the region of minimum variance is apparently slightly higher, around 3 kHz, again where the distinguishing spectral features of the sibilants occur. These results suggest that these regions may be a distinguishing feature of each of the sibilants, and this is most effectively demonstrated by figures 6.5 and 6.8.

However, it must be stressed that it is not our assertion that these measures relate to any 'classification' capability. Rather, this work attempts to investigate the existence of important characteristics within the variations of fricatives, and this is shown to be true.

The spectral shapes of the non-sibilants were found to be very similar, and overlapped significantly at every frequency in the spectrum. However, significant trends were observed. The frequency of the spectral peak in /f/ tokens usually occurs around 2 kHz, and if not here, would generally be found at either 4 kHz, or 7 kHz (for males) or 8 kHz (for females). However, for /θ/ tokens, the peak frequency is prone to much larger degree of variation.

Evidence has been found that the spectral variability of the non-sibilants over time may have different characteristics for /f/ and /θ/. A very crude measure of the total spectral variability over time reveals that male /θ/ tokens generally have a much lower degree of within-token spectral variation than /f/ tokens, although similar findings were not found for females.

An alternative method for attempting to capture that degree of variation over time was to track the spectral peak frequency through non-sibilant tokens. The variability of the spectral peak frequency also suggested that male /θ/ productions were much less variable than /f/ productions, though again, this could not be verified for the female productions.

Little additional information was provided by categorising spectral variance measures by vowel context. This finding may have implications concerning the effect of vowel context on fricative production.

Many potential techniques exist for capturing the subtle spectral variations that occur within fricative productions. Spectral covariance plots have been generated, but found little distinguishing information across the fricatives. Nevertheless, they provide intuitive appeal, and renewed efforts in this area may produce better results.

An analysis of spectral variations in fricative productions has revealed interesting new information, that is straightforward to interpret, and which agrees with the general theories of fricative production. While some of these results have provided useful distinguishing information, other results are less obvious, but reveal a significant amount about productions nonetheless.

### 8.1.6 Analysis of disordered speech

The new analysis techniques have been applied to a number of speech productions of subjects with cochlear implants.

In several cases these productions are clearly disordered, and this is often shown clearly in all of the spectral analysis techniques. In other less obvious cases, particularly productions with some slightly abnormal temporal 'fluctuation', the traditional methods (spectrograms and spectral moments) often do not suggest that the production is in any way abnormal. However, the new methods of spectral variance more commonly present evidence of abnormal productions.

These new techniques provide a diverse array of indispensable additional information, and it appears, additional important features of production. They are in particular better suited to dealing with productions containing unusual temporal features, since they are based upon assumptions that the fricatives are nonstationary.

## 8.2 Future Work

It has been shown that the production of fricatives is subject to physical processes that produce variations in different regions of the spectrum that depend upon the target fricative. It is through the analysis of these variations that a better understanding of fricative production will be obtained, and while the physical reasoning behind these is of ultimate interest, they are largely beyond the scope of this thesis. However, the results we have obtained are a strong foundation from which theorism and experimentation of the more intricate aspect of fricative production can begin.

Of course, many additional areas remain to be explored, using the methods developed here. While voiced fricatives were incorporated into the word corpus, they have not been analysed here. However, multitaper analysis should provide better spectral estimates of voiced fricatives, and in this regard, it is hoped that attempts to describe the voiced fricatives will advance on previous efforts. Spectral moments have elsewhere been applied to the voiced fricatives, although their use is highly limited, since a significant characteristic of the spectral shape of the voiced fricatives is the spectral peaks due to voicing, and such features are known to not be well captured by spectral moments.

It also seems necessary to place more emphasis on the boundary regions of the fricatives, since it seems more likely than ever that information pertaining to the non-sibilants is to be found here. The data collected for this thesis are suitable for the analysis of the boundary region. However, this was not attempted, since observation of the multitaper spectrograms provided no clear approach along which to proceed. It is possible that such analysis may be able to make use of the spectral covariance methods investigated.

These new methods are to be used in a more thorough analysis of cochlear implant subjects. Changes over time of the productions of such subjects should be more readily interpretable from these more intuitive analysis techniques, and problems concerning nonstationary elements are much less likely to be encountered.

New opportunities have been opened up for analysis of fricative productions. The search for acoustical fricative production characteristics continues, and promises to be most rewarding.

# Appendix A

# Corpus

The following words were used to capture the fricatives in the desired vowel contexts.

| Context | /f/ | /θ/ | /s/ | /ʃ/ |
|---|---|---|---|---|
| /iFi/ | 'beefy' | 'teethy' | 'fleecy' | 'quichey' |
| /ˈiFə/ | 'beefer' | 'ether' | 'Lisa' | 'Letitia' |
| /ˈuFi/ | 'goofy' | 'toothy' | 'Lucy' | 'sushi' |
| /ˈuFə/ | 'loofah' | 'Luther' | 'juicer' | 'fuchsia' |
| /əFˈi/ | 'atrophy' | 'Athena' | 'casino' | 'machine' |
| /əFˈu/ | 'buffoon' | 'Methuselah' | 'bassoon' | 'parachute' |

| Context | /v/ | /ð/ | /z/ | /ʒ/ |
|---|---|---|---|---|
| /iFi/ | 'D.V.D.' | 'I will see thee' | 'easy' | 'Gigi' |
| /ˈiFə/ | 'leaver' | 'breather' | 'teaser' | 'seizure' |
| /ˈuFi/ | 'groovy' | 'smoothie' | 'boozy' | 'bijou'[1] |
| /ˈuFə/ | 'Hoover' | 'smoother' | 'cruiser' | 'Hoosier' |
| /əFˈi/ | 'Davina' | 'I sing to thee' | 'magazine' | 'regime' |
| /əFˈu/ | 'the voodoo-doll' | 'give Eva those' | 'bazooka' | 'jejune' |

Each page of words, or 'script' to be read by the speaker is now presented. Each 'script' is separated by a horizontal line. The first 'script', on each of the following pages, was used as a test page only.

[1] Reversed vowel order.

| | | | |
|---|---|---|---|
| “fleecy” | “teethy” | “quichey” | “beefy” |
| “Lisa” | “ether” | “Letitia” | “beefer” |
| “Lucy” | “toothy” | “sushi” | “goofy” |
| “juicer” | “Luther” | “fuchsia” | “loofah” |
| “casino” | “Athena” | “machine” | “atrophy” |
| “bassoon” | “Methuselah” | “parachute” | “buffoon” |
| “beefy” | “teethy” | “fleecy” | “quichey” |
| “beefer” | “ether” | “Lisa” | “Letitia” |
| “goofy” | “toothy” | “Lucy” | “sushi” |
| “loofah” | “Luther” | “juicer” | “fuchsia” |
| “atrophy” | “Athena” | “casino” | “machine” |
| “buffoon” | “Methuselah” | “bassoon” | “parachute” |
| “teethy” | “fleecy” | “quichey” | “beefy” |
| “ether” | “Lisa” | “Letitia” | “beefer” |
| “toothy” | “Lucy” | “sushi” | “goofy” |
| “Luther” | “juicer” | “fuchsia” | “loofah” |
| “Athena” | “casino” | “machine” | “atrophy” |
| “Methuselah” | “bassoon” | “parachute” | “buffoon” |
| “fleecy” | “quichey” | “beefy” | “teethy” |
| “Lisa” | “Letitia” | “beefer” | “ether” |
| “Lucy” | “sushi” | “goofy” | “toothy” |
| “juicer” | “fuchsia” | “loofah” | “Luther” |
| “casino” | “machine” | “atrophy” | “Athena” |
| “bassoon” | “parachute” | “buffoon” | “Methuselah” |
| “quichey” | “beefy” | “teethy” | “fleecy” |
| “Letitia” | “beefer” | “ether” | “Lisa” |
| “sushi” | “goofy” | “toothy” | “Lucy” |
| “fuchsia” | “loofah” | “Luther” | “juicer” |
| “machine” | “atrophy” | “Athena” | “casino” |
| “parachute” | “buffoon” | “Methuselah” | “bassoon” |
| “beefy” | “fleecy” | “teethy” | “quichey” |
| “beefer” | “Lisa” | “ether” | “Letitia” |
| “goofy” | “Lucy” | “toothy” | “sushi” |
| “loofah” | “juicer” | “Luther” | “fuchsia” |
| “atrophy” | “casino” | “Athena” | “machine” |
| “buffoon” | “bassoon” | “Methuselah” | “parachute” |
| “fleecy” | “beefy” | “quichey” | “teethy” |
| “Lisa” | “beefer” | “Letitia” | “ether” |
| “Lucy” | “goofy” | “sushi” | “toothy” |
| “juicer” | “loofah” | “fuchsia” | “Luther” |
| “casino” | “atrophy” | “machine” | “Athena” |
| “bassoon” | “buffoon” | “parachute” | “Methuselah” |

| | | | |
|---|---|---|---|
| “Gigi” | “easy” | “D.V.D.” | “seething” |
| “seizure” | “teaser” | “leaver” | “breather” |
| “Hoosier” | “cruiser” | “Hoover” | “smoother” |
| “regime” | “magazine” | “Davina” | “I sing to thee.” |
| “groovy” | “boozy” | “smoothie” | |
| “the voodoo-doll” | “I will see thee” | “give Eva those” | |
| “D.V.D.” | “seething” | “easy” | “Gigi” |
| “leaver” | “breather” | “teaser” | “seizure” |
| “Hoover” | “smoother” | “cruiser” | “Hoosier” |
| “Davina” | “I sing to thee.” | “magazine” | “regime” |
| “groovy” | “smoothie” | “boozy” | |
| “the voodoo-doll” | “give Eva those” | “I will see thee” | |
| “seething” | “easy” | “Gigi” | “D.V.D.” |
| “breather” | “teaser” | “seizure” | “leaver” |
| “smoother” | “cruiser” | “Hoosier” | “Hoover” |
| “I sing to thee.” | “magazine” | “regime” | “Davina” |
| “smoothie” | “boozy” | “groovy” | |
| “give Eva those” | “I will see thee” | “the voodoo-doll” | |
| “easy” | “Gigi” | “D.V.D.” | “seething” |
| “teaser” | “seizure” | “leaver” | “breather” |
| “cruiser” | “Hoosier” | “Hoover” | “smoother” |
| “magazine” | “regime” | “Davina” | “I sing to thee.” |
| “boozy” | “groovy” | “smoothie” | |
| “I will see thee” | “the voodoo-doll” | “give Eva those” | |
| “Gigi” | “D.V.D.” | “seething” | “easy” |
| “seizure” | “leaver” | “breather” | “teaser” |
| “Hoosier” | “Hoover” | “smoother” | “cruiser” |
| “regime” | “Davina” | “I sing to thee.” | “magazine” |
| “groovy” | “smoothie” | “boozy” | |
| “the voodoo-doll” | “give Eva those” | “I will see thee” | |
| “D.V.D.” | “easy” | “seething” | “Gigi” |
| “leaver” | “teaser” | “breather” | “seizure” |
| “Hoover” | “cruiser” | “smoother” | “Hoosier” |
| “Davina” | “magazine” | “I sing to thee.” | “regime” |
| “groovy” | “boozy” | “smoothie” | |
| “the voodoo-doll” | “I will see thee” | “give Eva those” | |
| “Gigi” | “seething” | “easy” | “D.V.D.” |
| “seizure” | “breather” | “teaser” | “leaver” |
| “Hoosier” | “smoother” | “cruiser” | “Hoover” |
| “regime” | “I sing to thee.” | “magazine” | “Davina” |
| “smoothie” | “boozy” | “groovy” | |
| “give Eva those” | “I will see thee” | “the voodoo-doll” | |

# Appendix B

# Multitaper Spectrograms

## B.1 Males

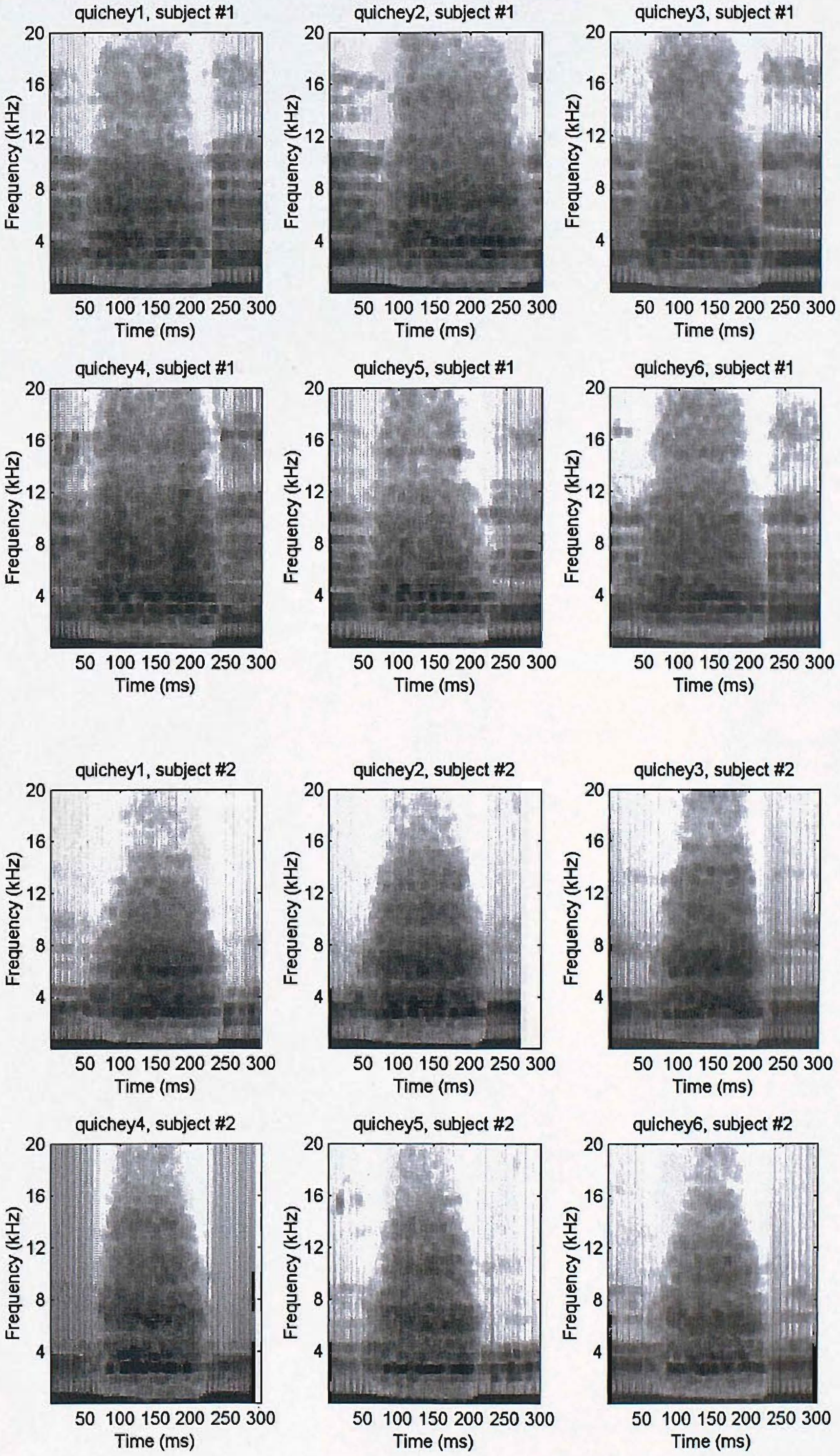


FIGURE B.1: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects M-01 and M-02.

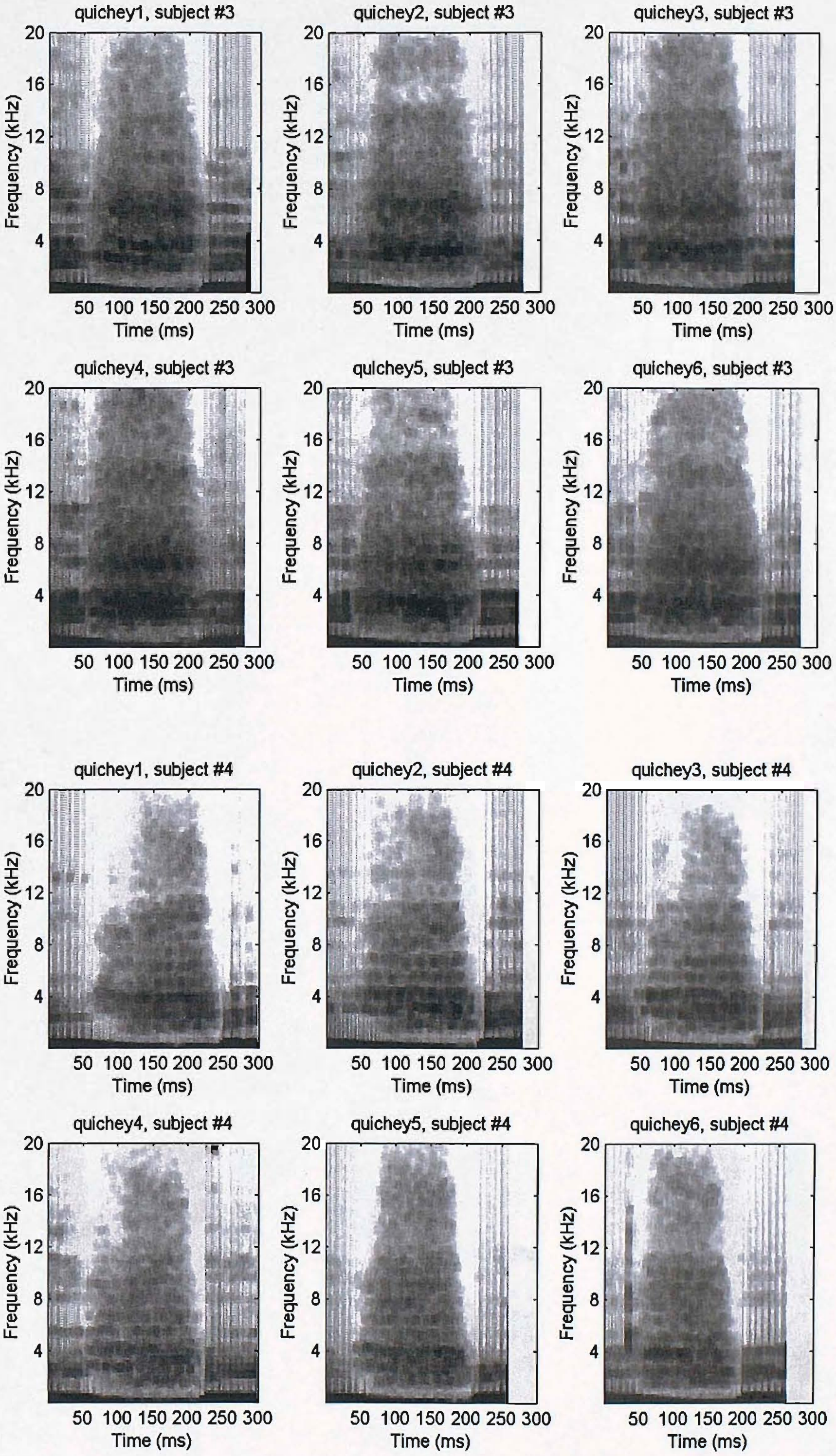


FIGURE B.2: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects M-03 and M-04.

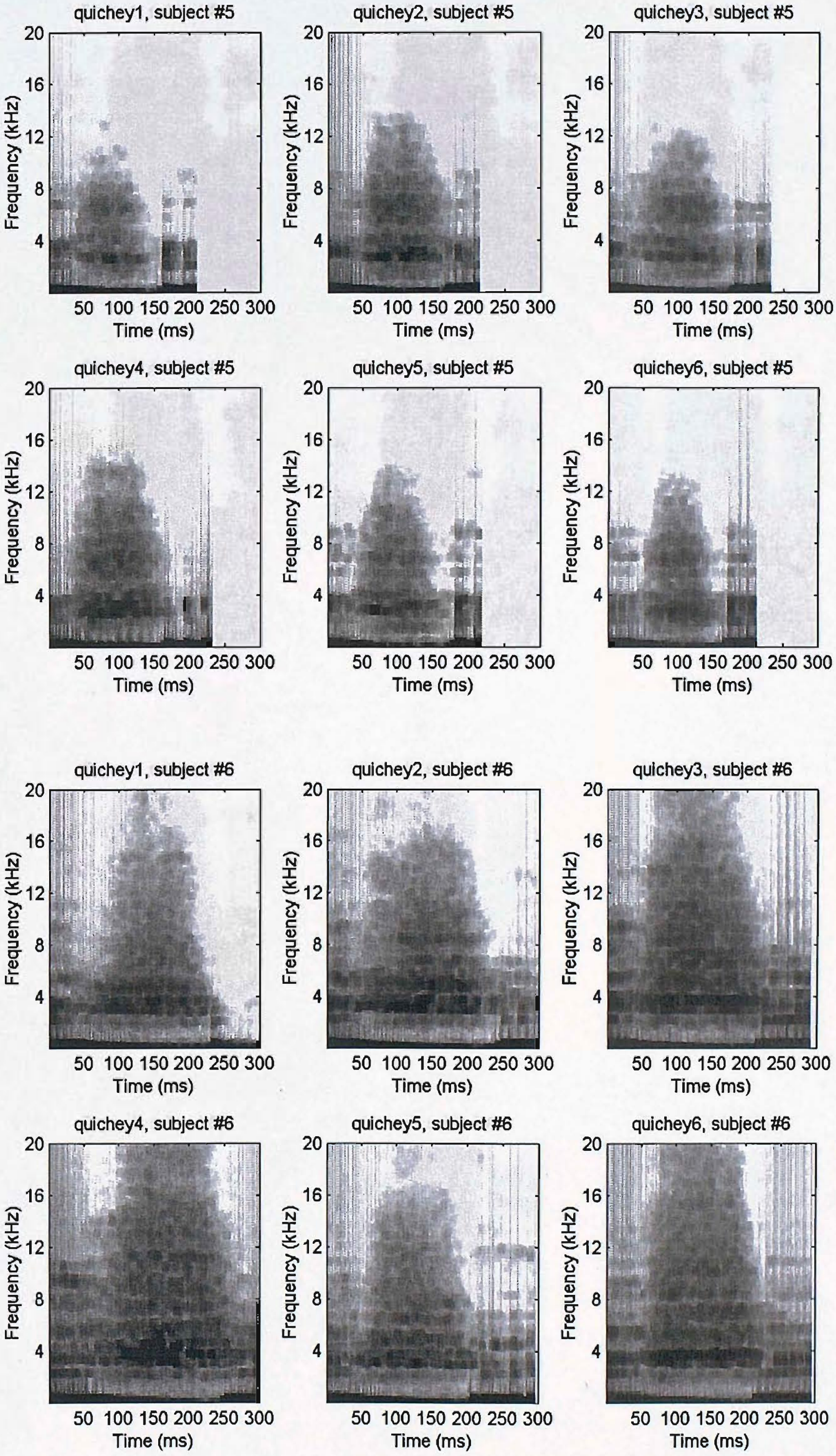


FIGURE B.3: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects M-05 and M-06.

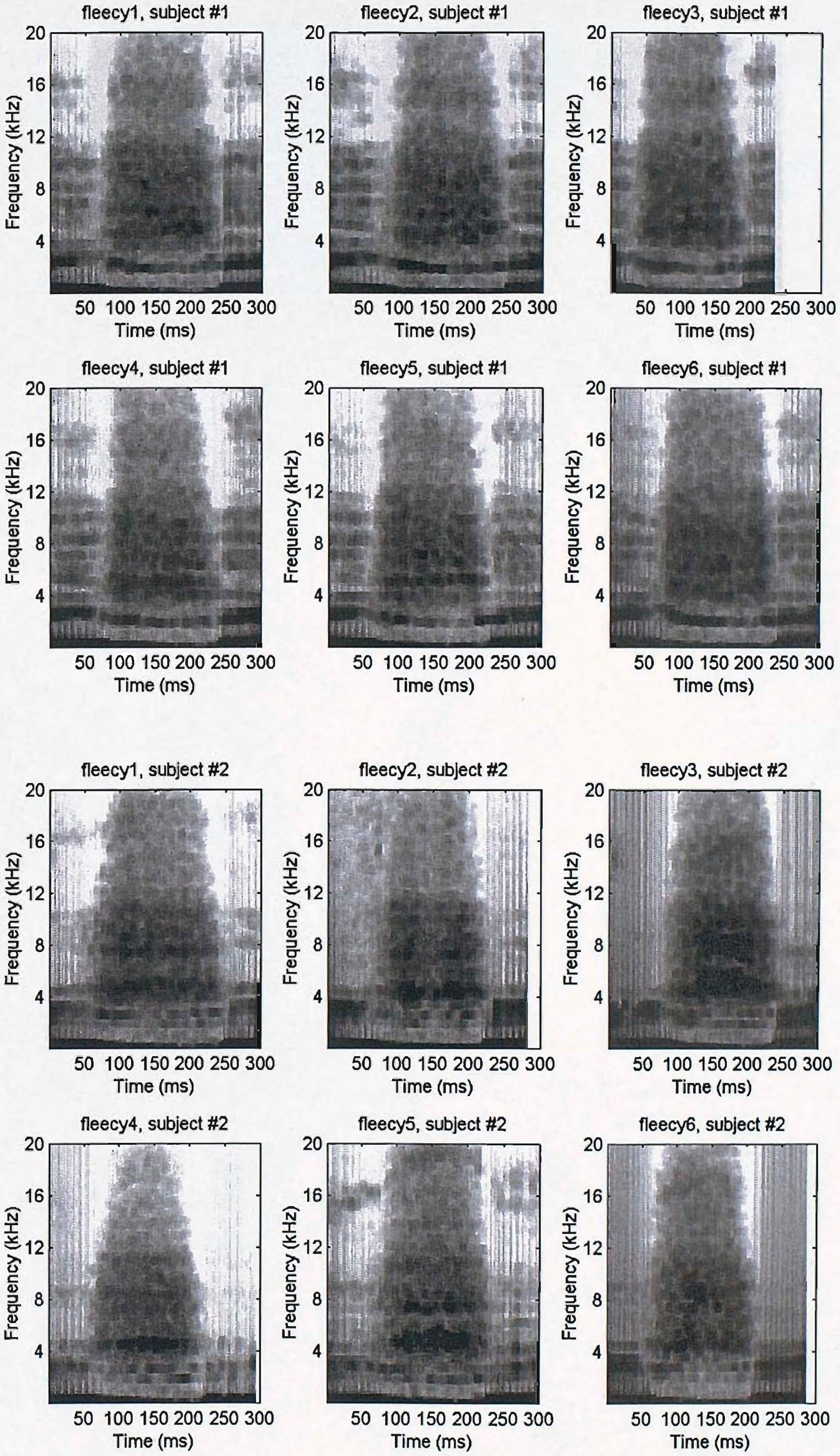


FIGURE B.4: Multitaper spectrograms: [ˈisi] productions from "fleecy", subjects M-01 and M-02.

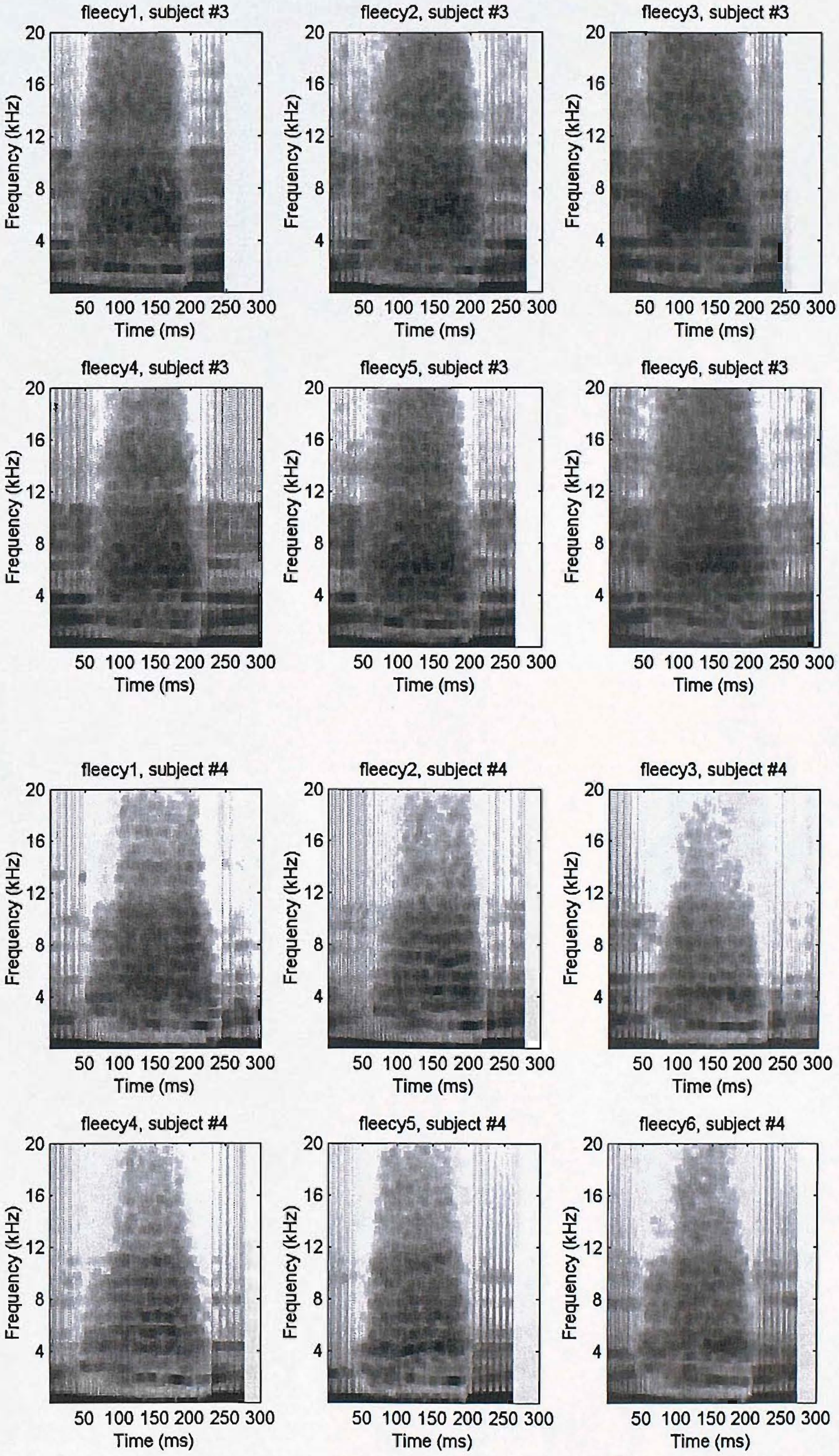


FIGURE B.5: Multitaper spectrograms: ['isi] productions from "fleecy", subjects M-03 and M-04.

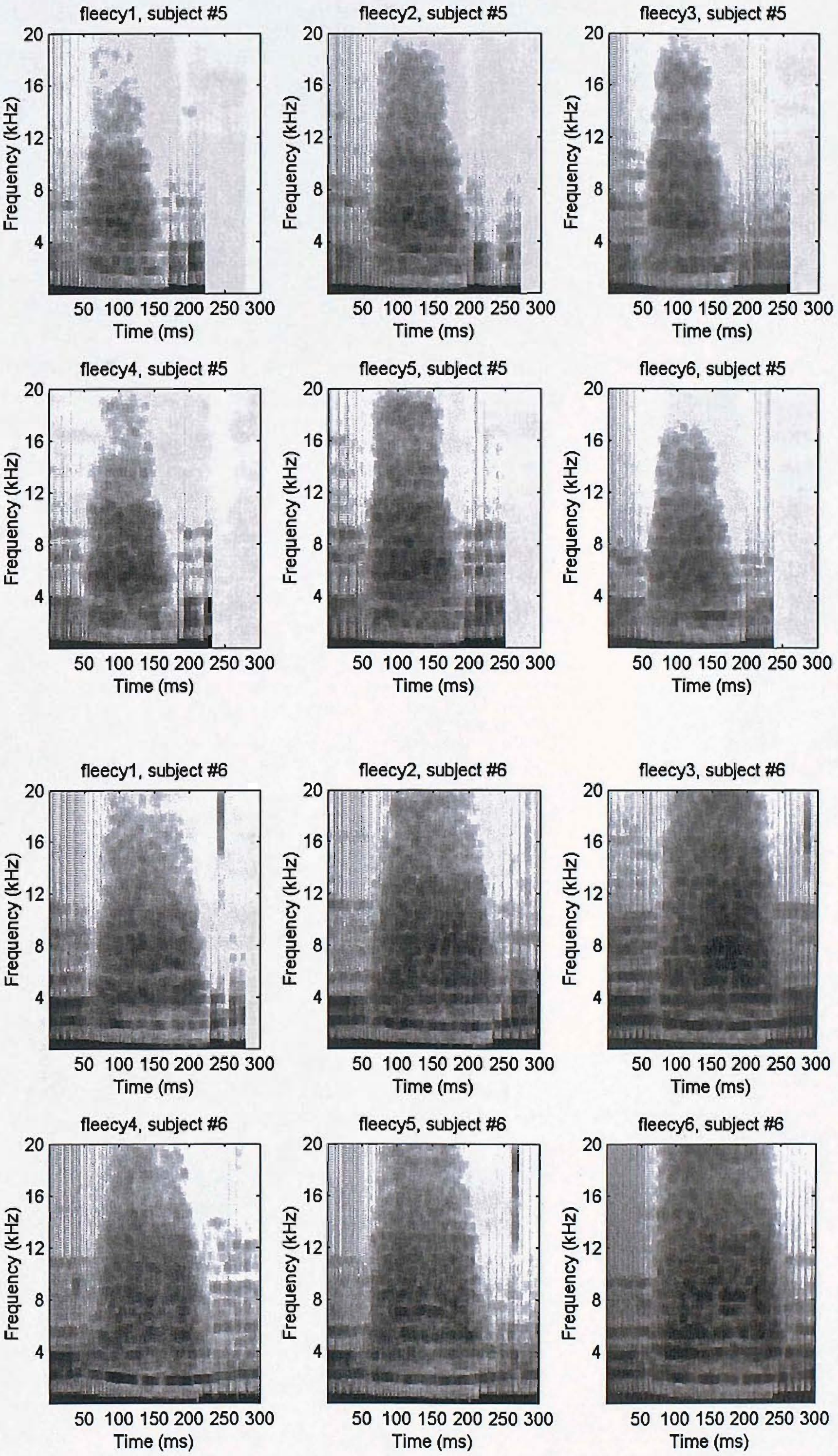
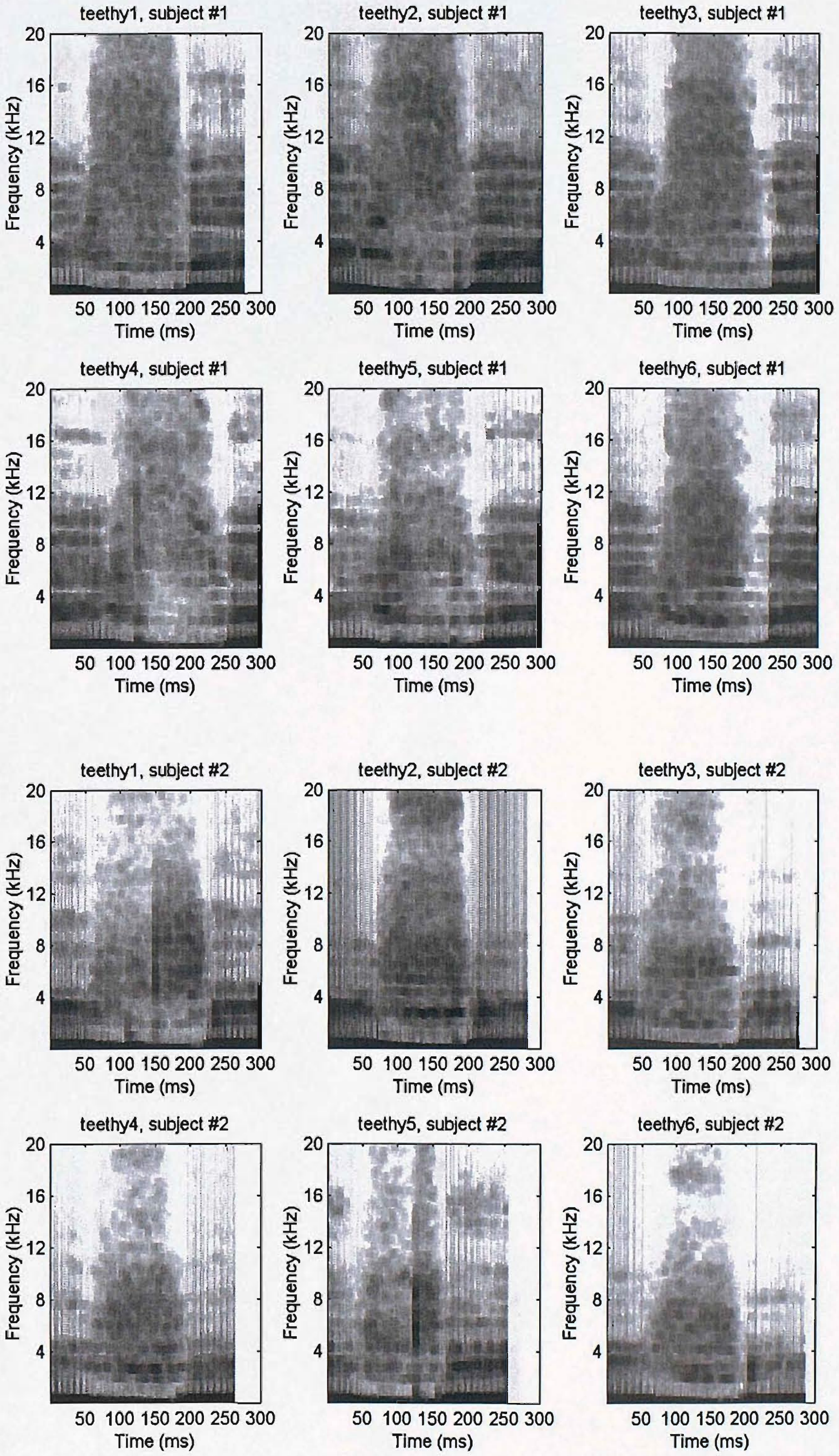


FIGURE B.6: Multitaper spectrograms: [ˈisi] productions from "fleecy", subjects M-05 and M-06.



teethy1, subject #1

teethy2, subject #1

teethy3, subject #1

teethy4, subject #1

teethy5, subject #1

teethy6, subject #1

teethy1, subject #2

teethy2, subject #2

teethy3, subject #2

teethy4, subject #2

teethy5, subject #2

teethy6, subject #2

FIGURE B.7: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects M-01 and M-02.

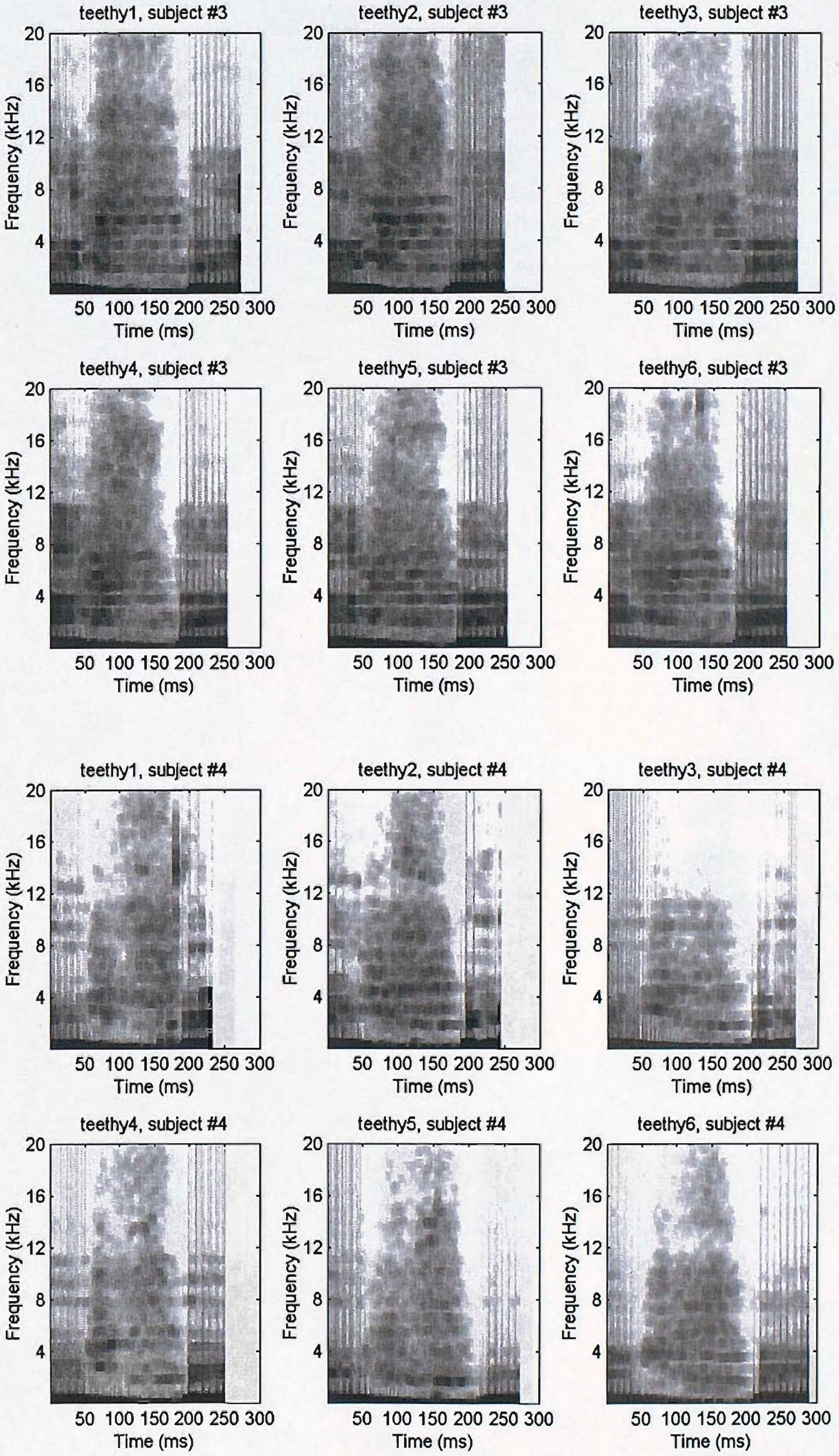


FIGURE B.8: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects M-03 and M-04.

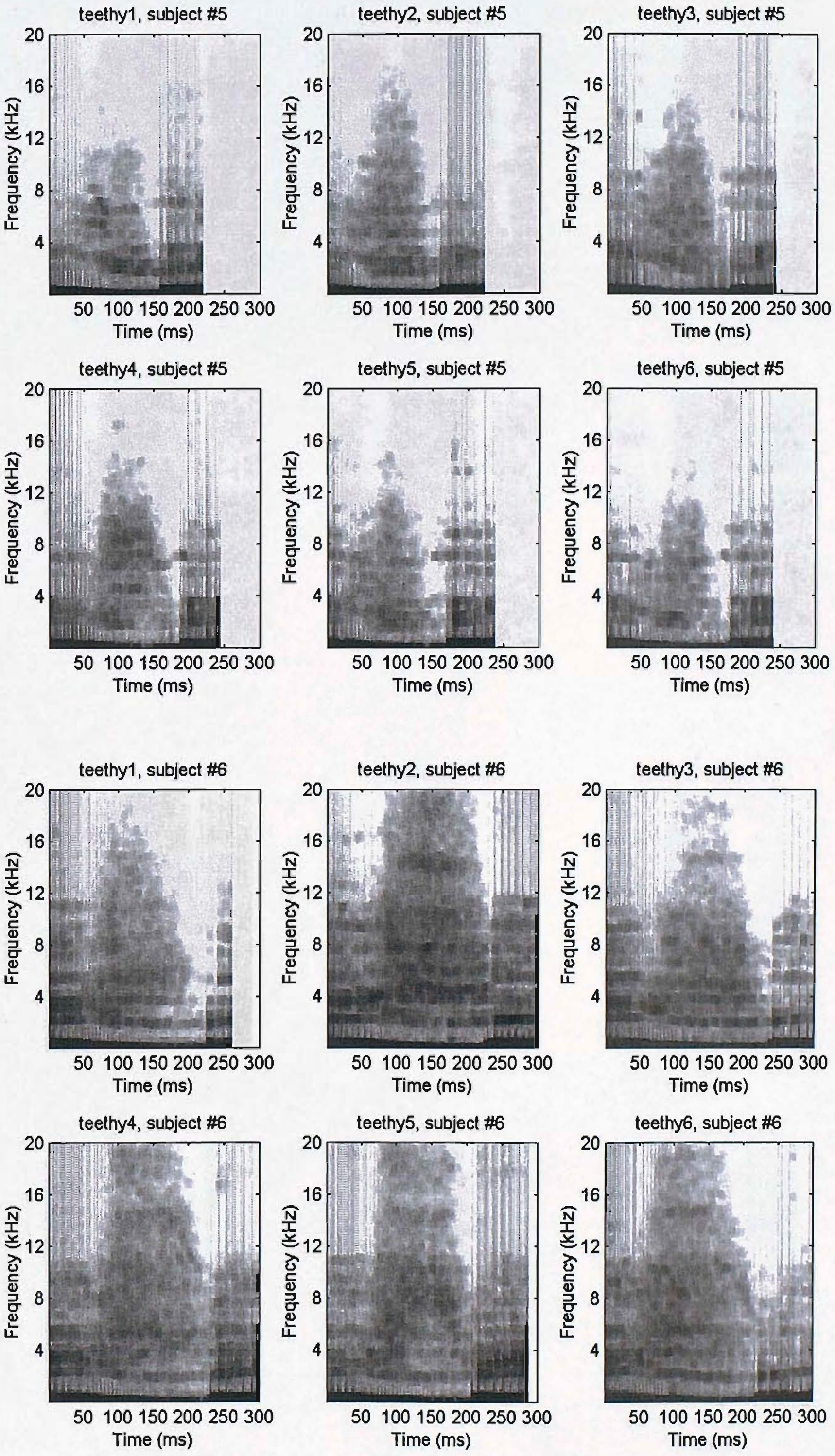


FIGURE B.9: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects M-05 and M-06.

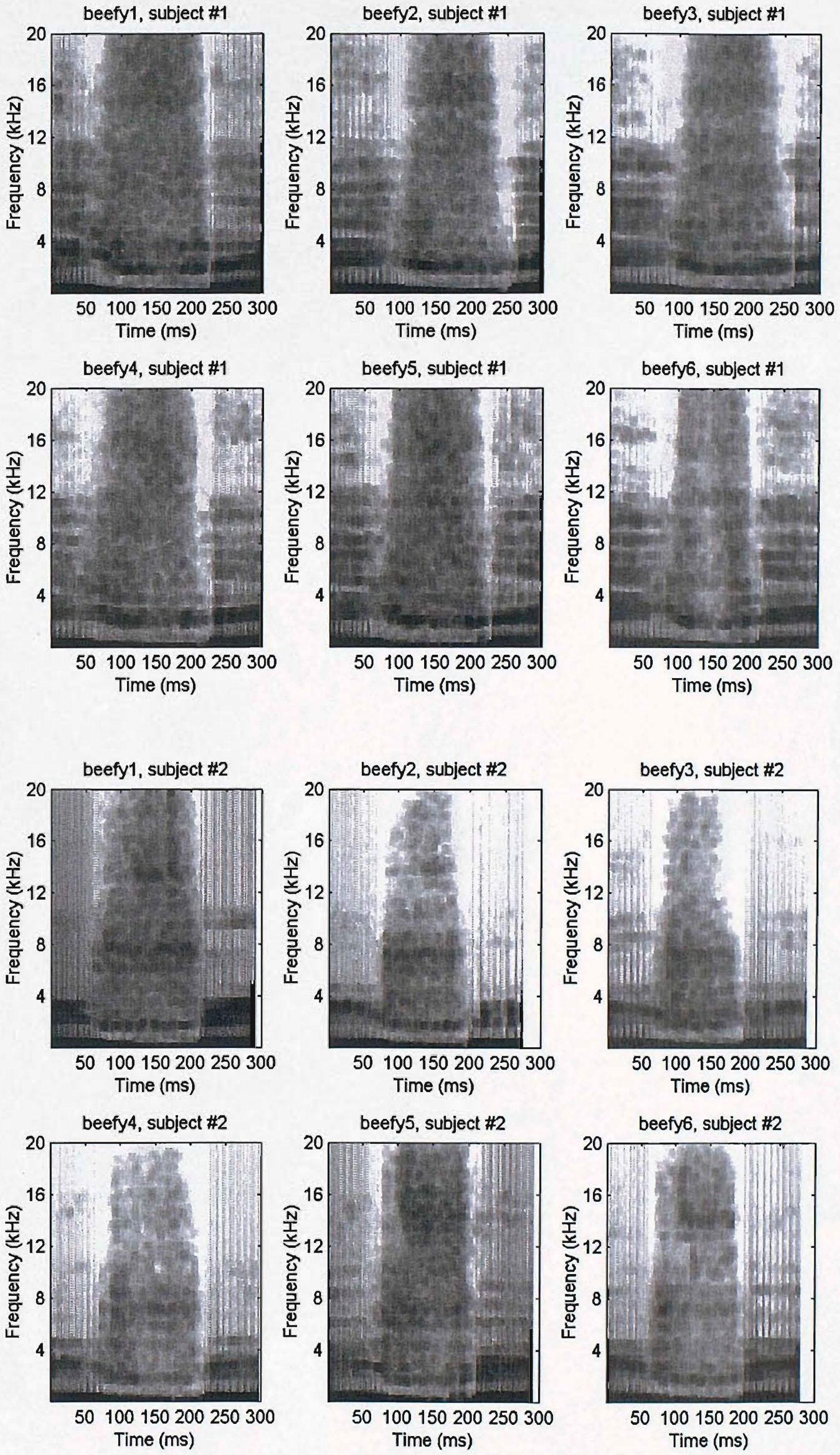
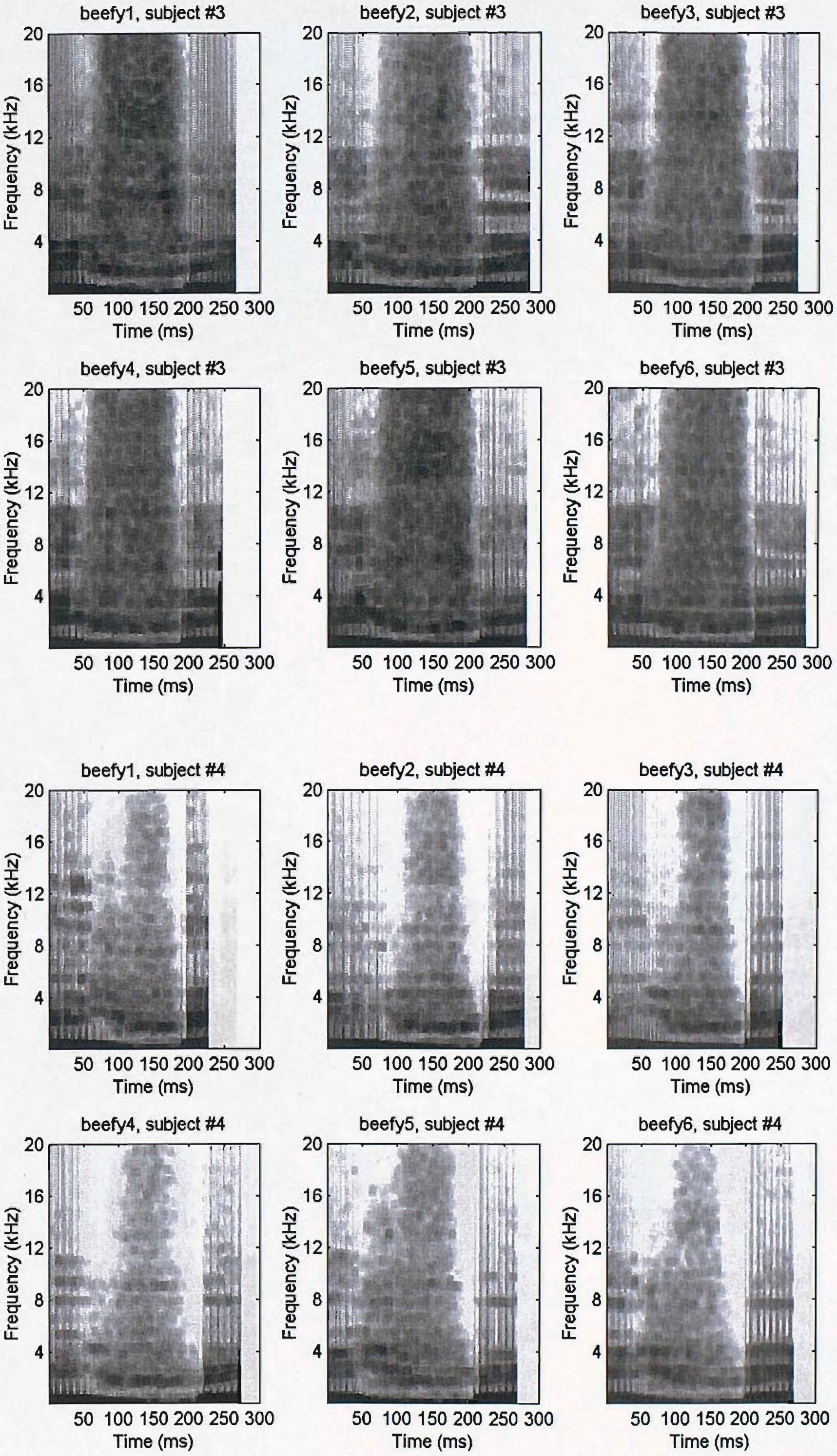


FIGURE B.10: Multitaper spectrograms: [ˈifi] productions from "beefy", subjects M-01 and M-02.



beefy1, subject #3

beefy2, subject #3

beefy3, subject #3

beefy4, subject #3

beefy5, subject #3

beefy6, subject #3

beefy1, subject #4

beefy2, subject #4

beefy3, subject #4

beefy4, subject #4

beefy5, subject #4

beefy6, subject #4

FIGURE B.11: Multitaper spectrograms: [ˈifi] productions from "beefy", subjects M-03 and M-04.

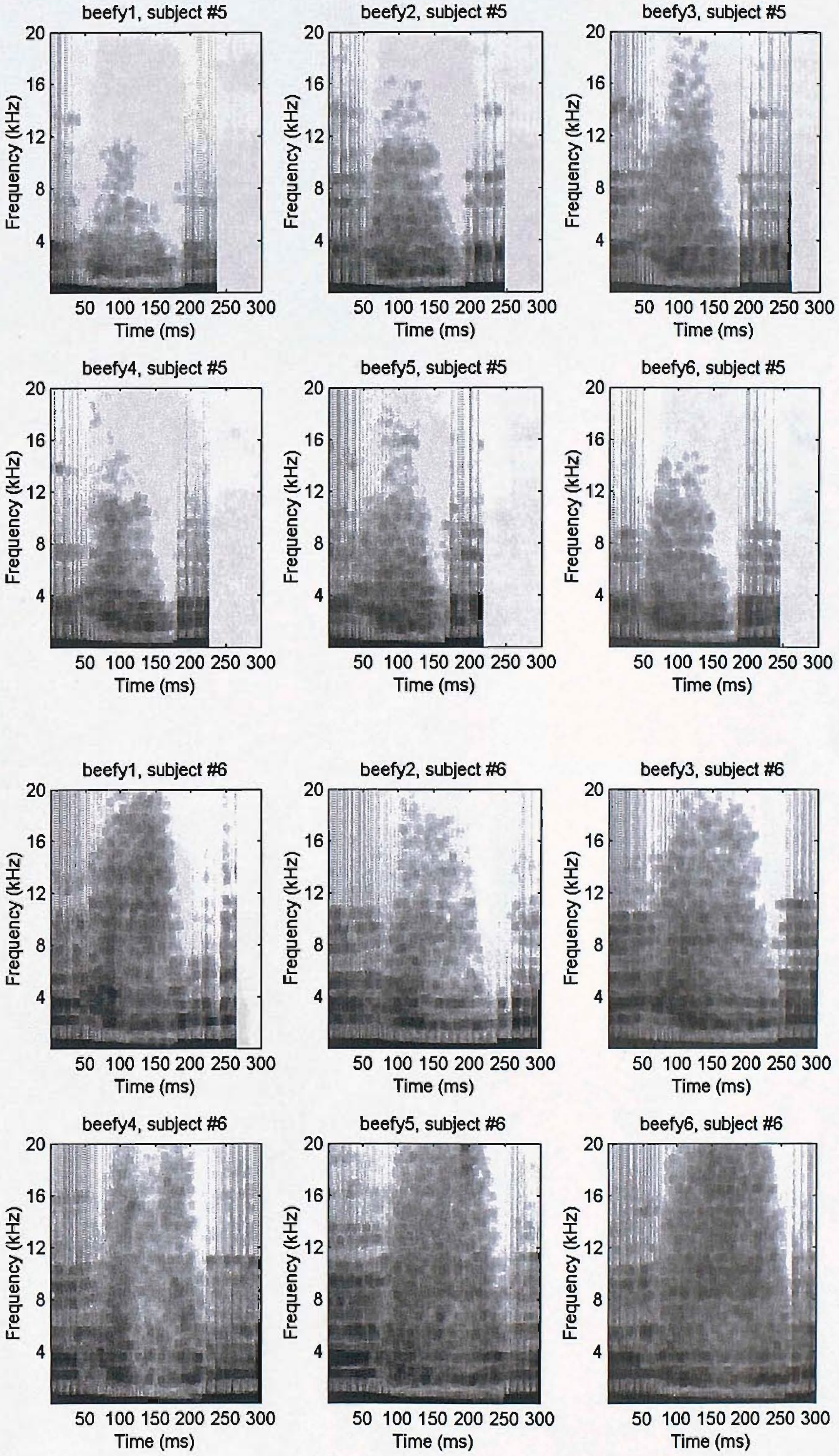


FIGURE B.12: Multitaper spectrograms: [ˈifi] productions from "beefy", subjects M-05 and M-06.

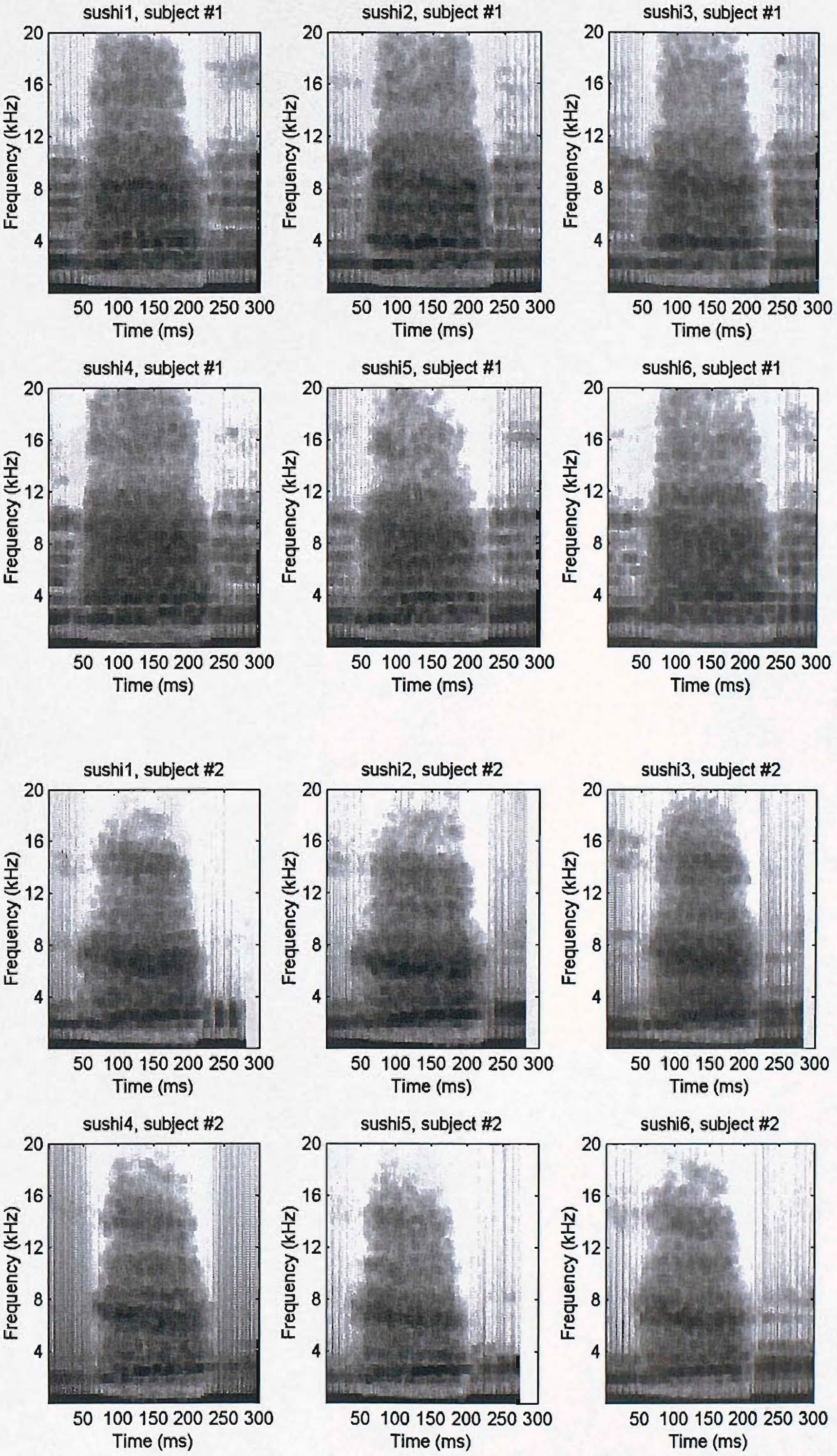


FIGURE B.13: Multitaper spectrograms: [ˈuʃi] productions from "sushi, subjects M-01 and M-02.

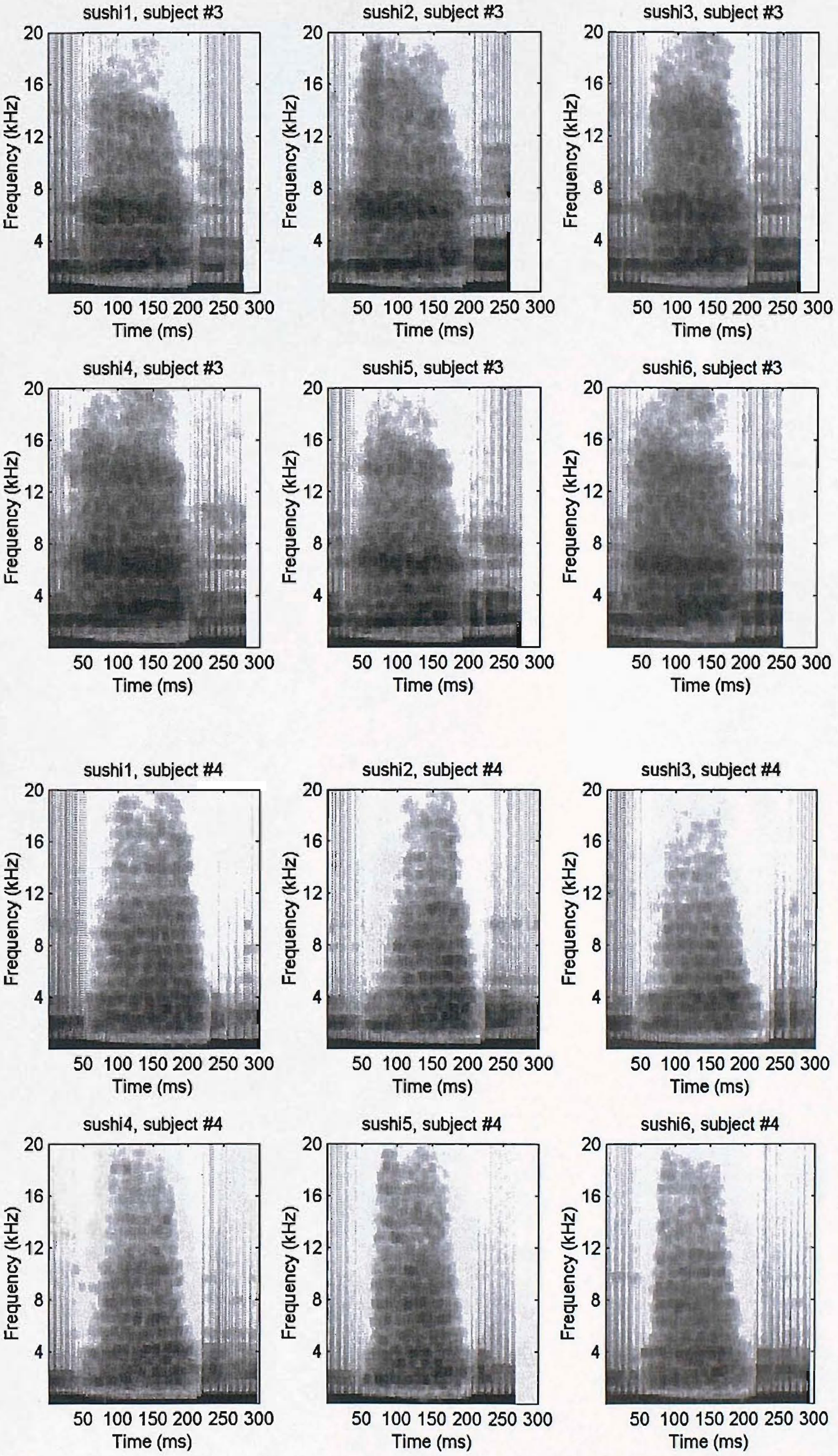


FIGURE B.14: Multitaper spectrograms: [ˈuʃi] productions from "sushi, subjects M-03 and M-04.

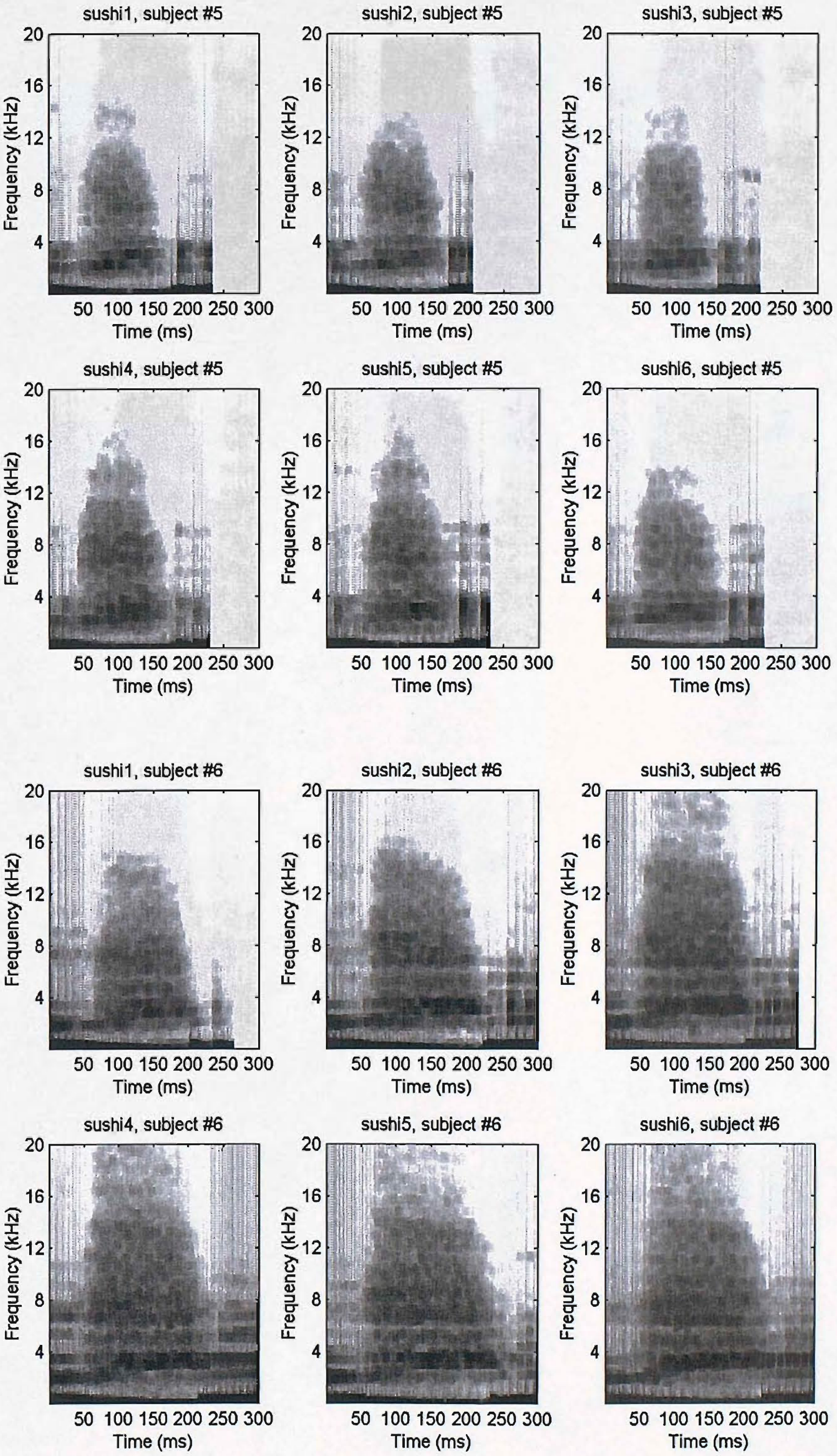


FIGURE B.15: Multitaper spectrograms: ['uʃi] productions from “sushi, subjects M-05 and M-06.

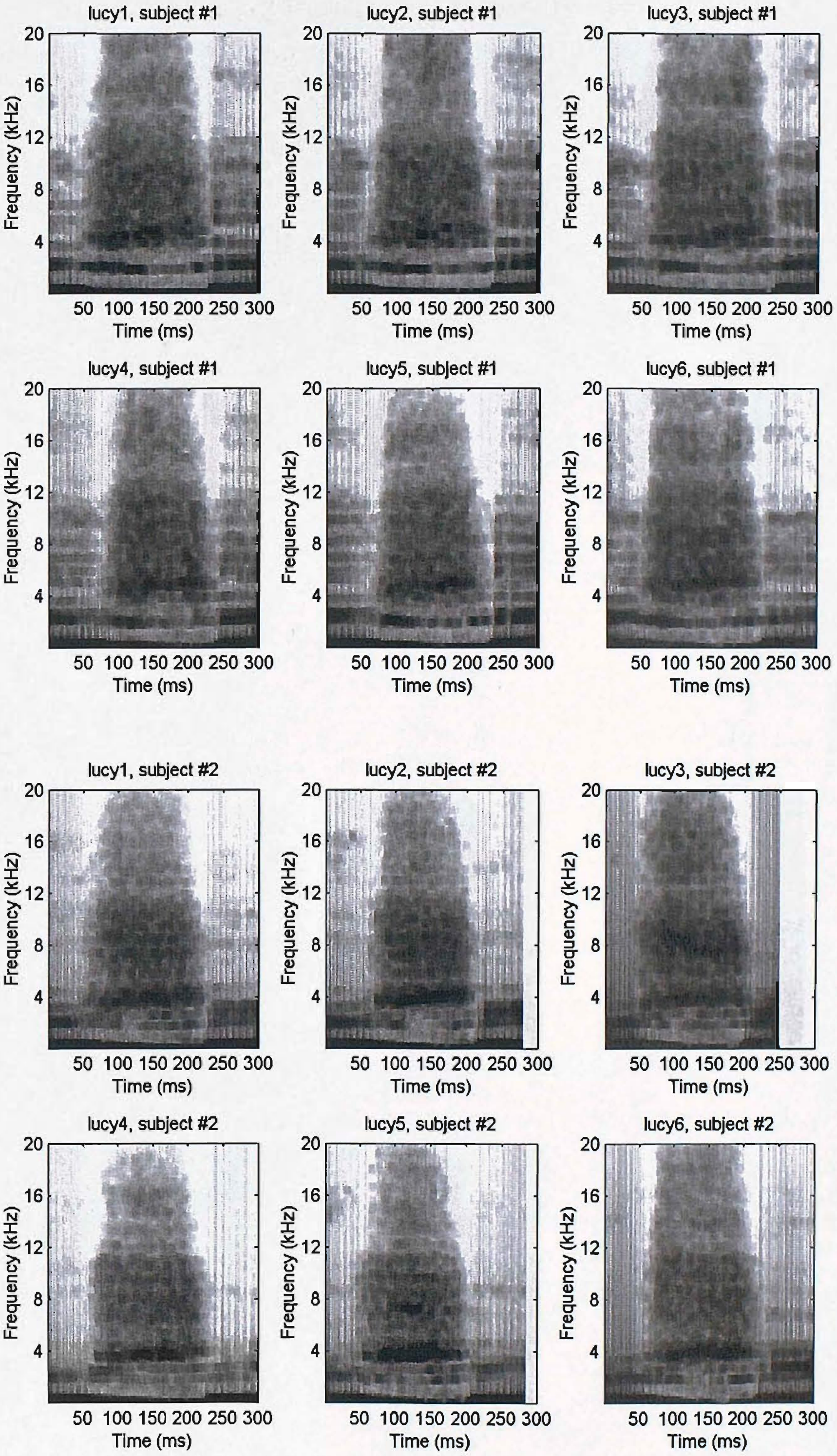


FIGURE B.16: Multitaper spectrograms: ['usi] productions from "Lucy", subjects M-01 and M-02.

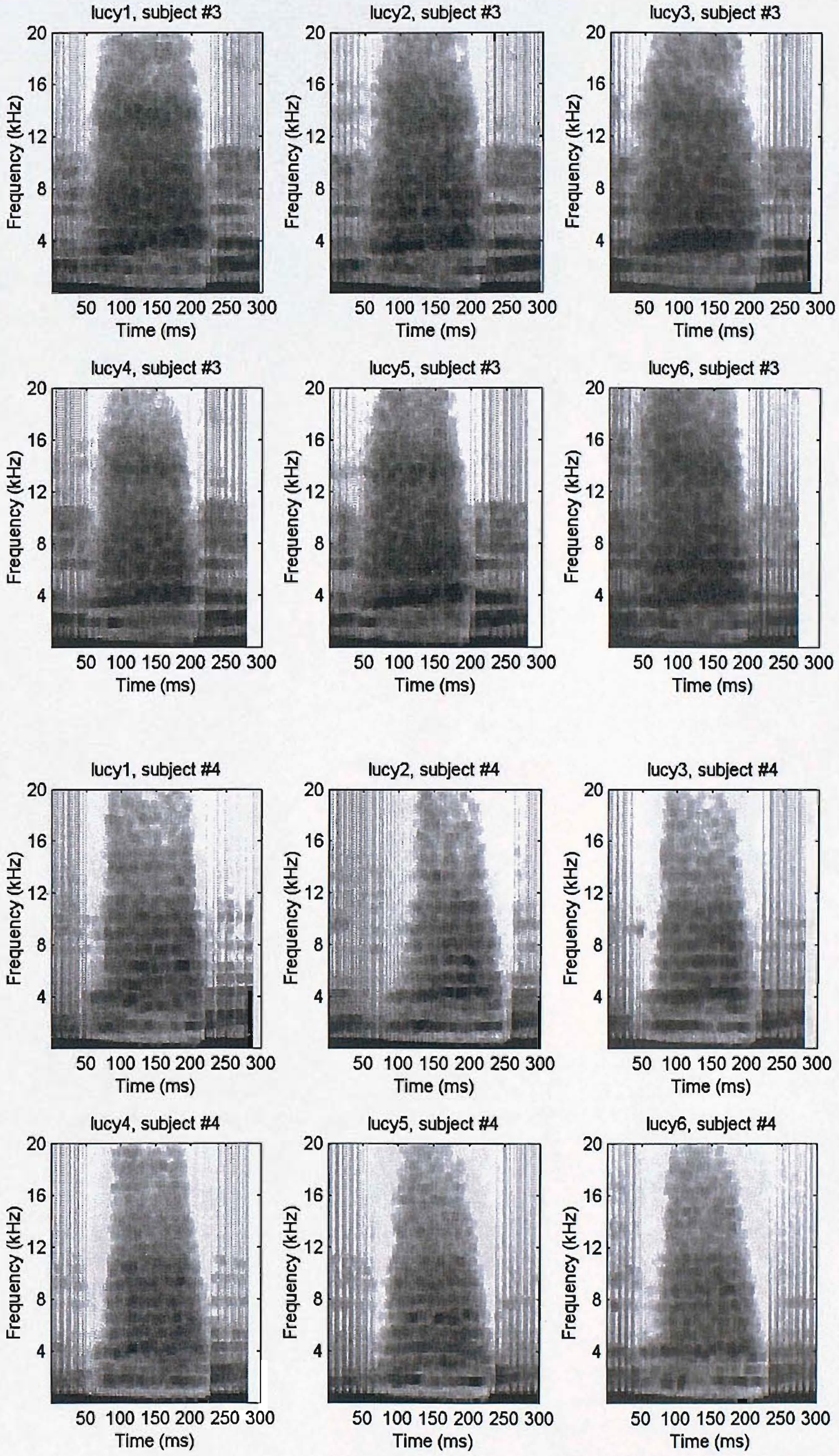


FIGURE B.17: Multitaper spectrograms: [ˈusi] productions from "Lucy", subjects M-03 and M-04.

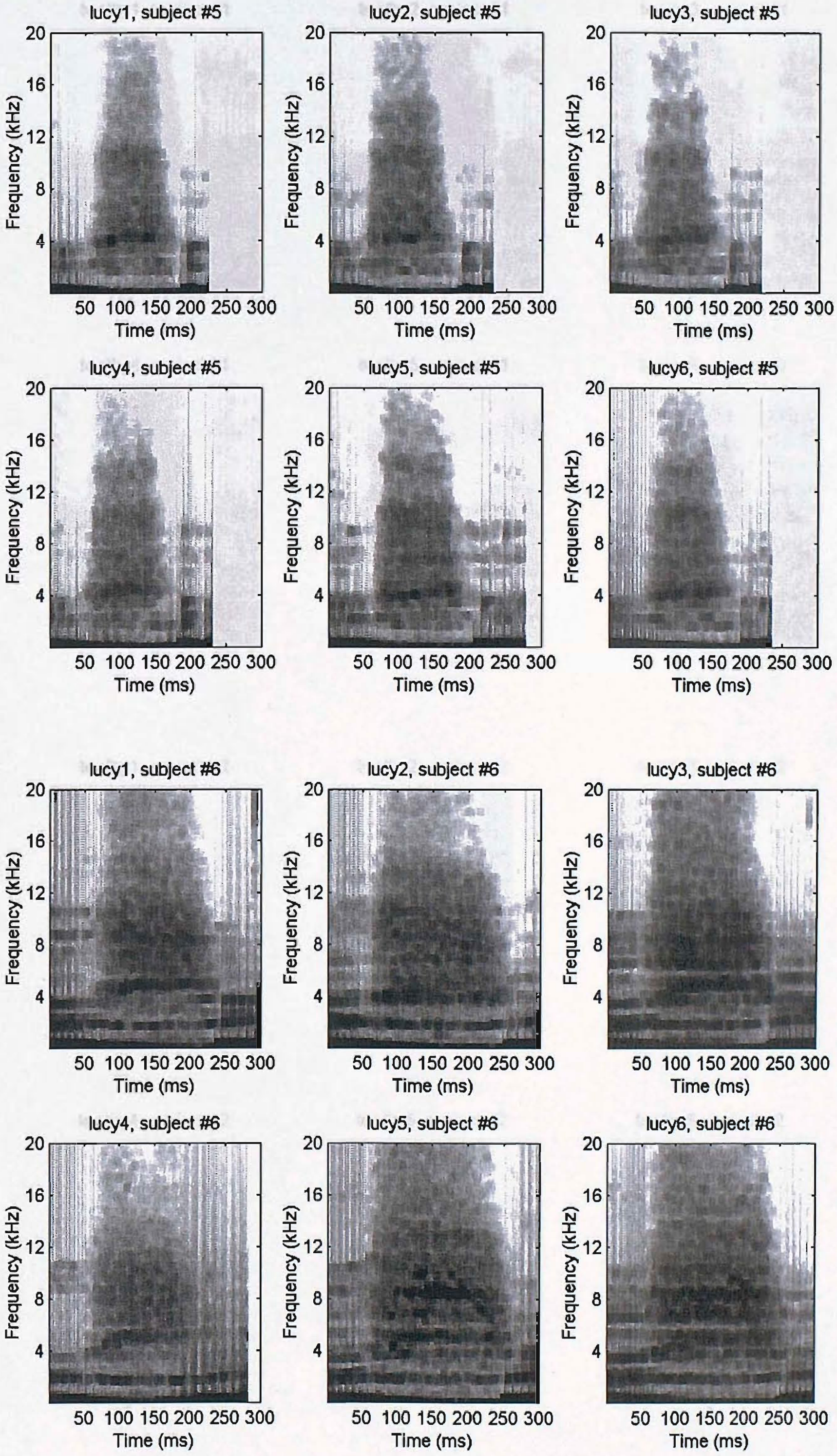


FIGURE B.18: Multitaper spectrograms: ['usi] productions from "Lucy", subjects M-05 and M-06.

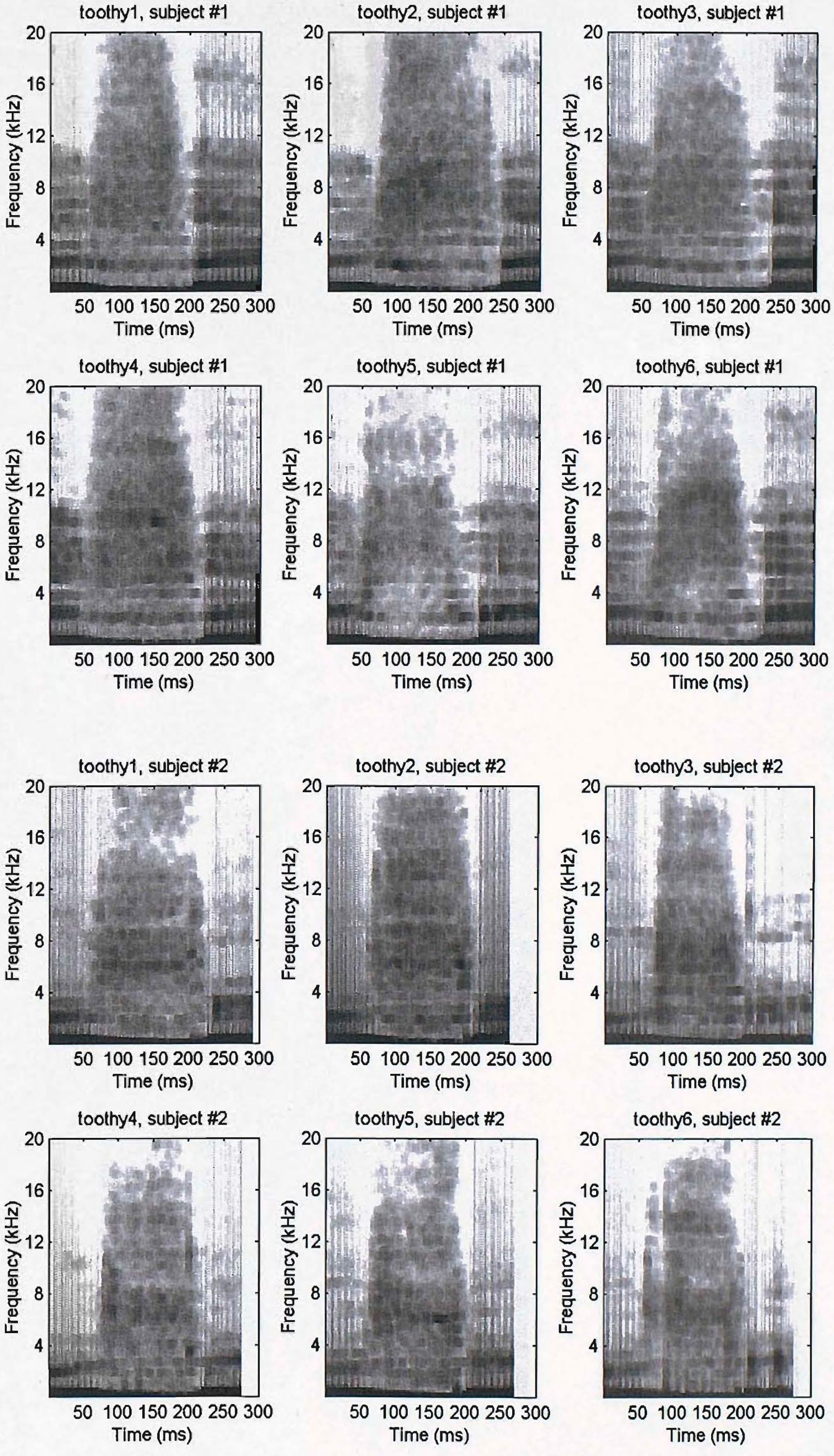


FIGURE B.19: Multitaper spectrograms: [ˈuθi] productions from "toothy", subjects M-01 and M-02.

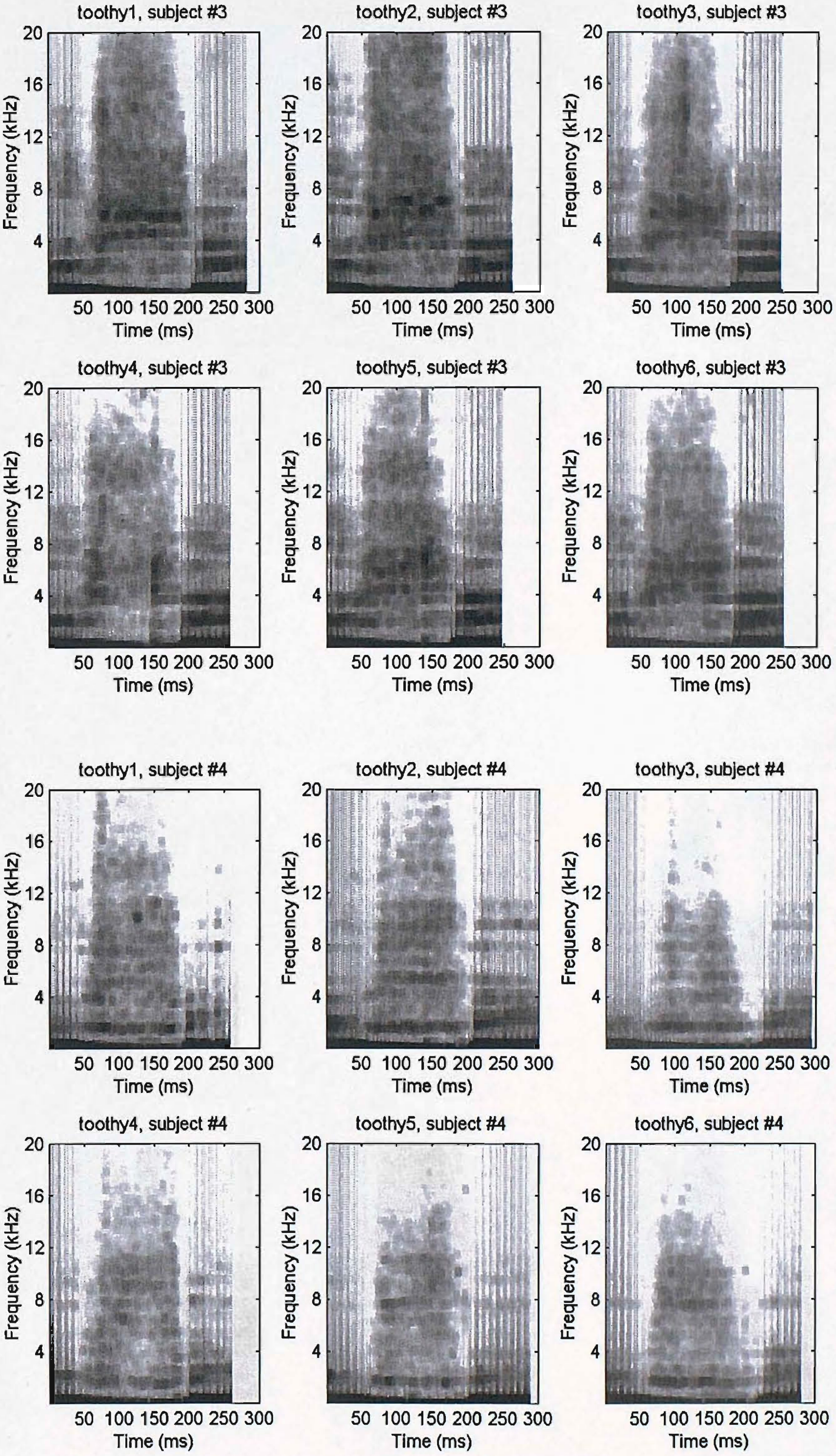


FIGURE B.20: Multitaper spectrograms: [ˈuθi] productions from "toothy", subjects M-03 and M-04.

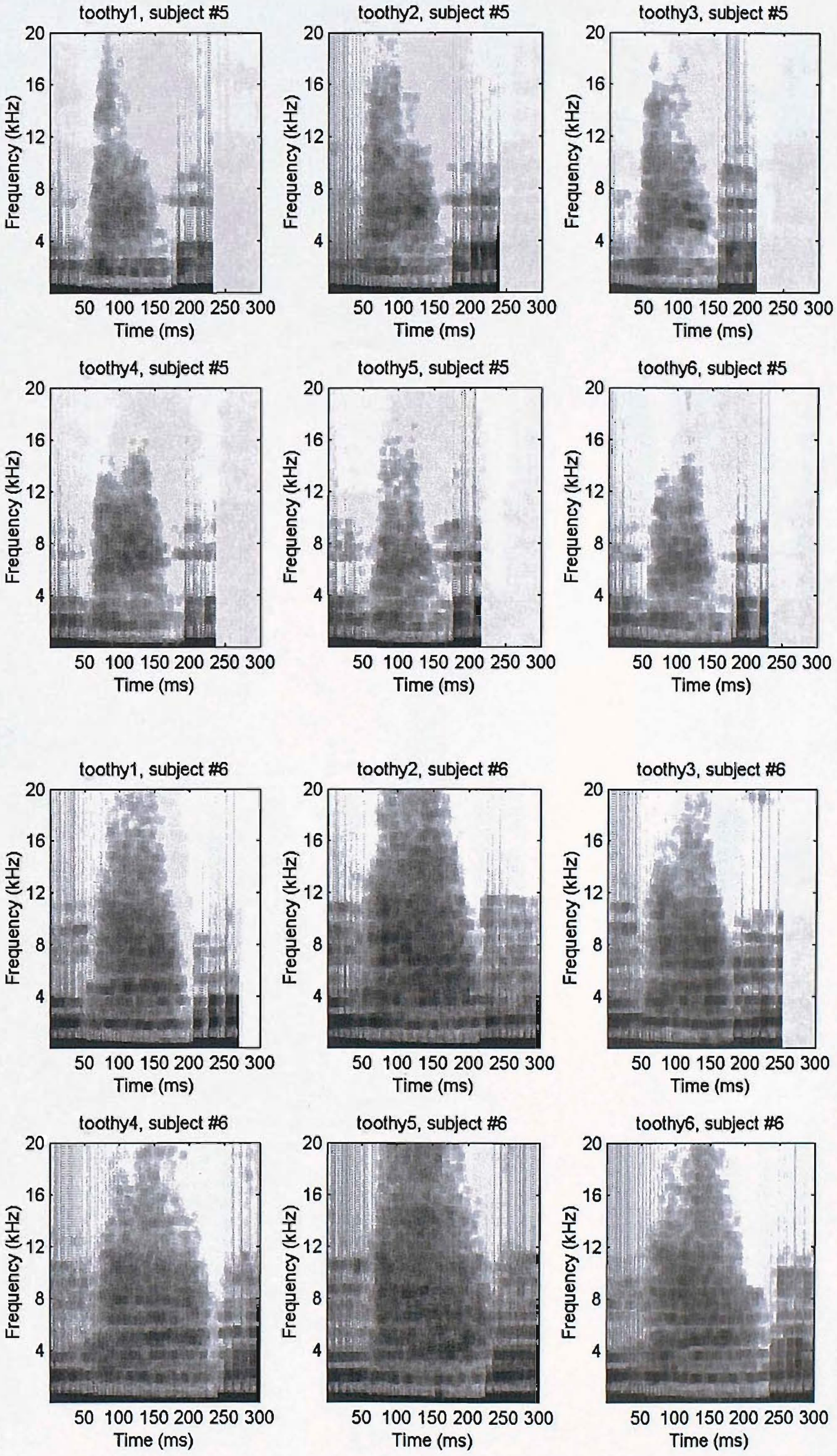


FIGURE B.21: Multitaper spectrograms: ['uθi] productions from "toothy", subjects M-05 and M-06.

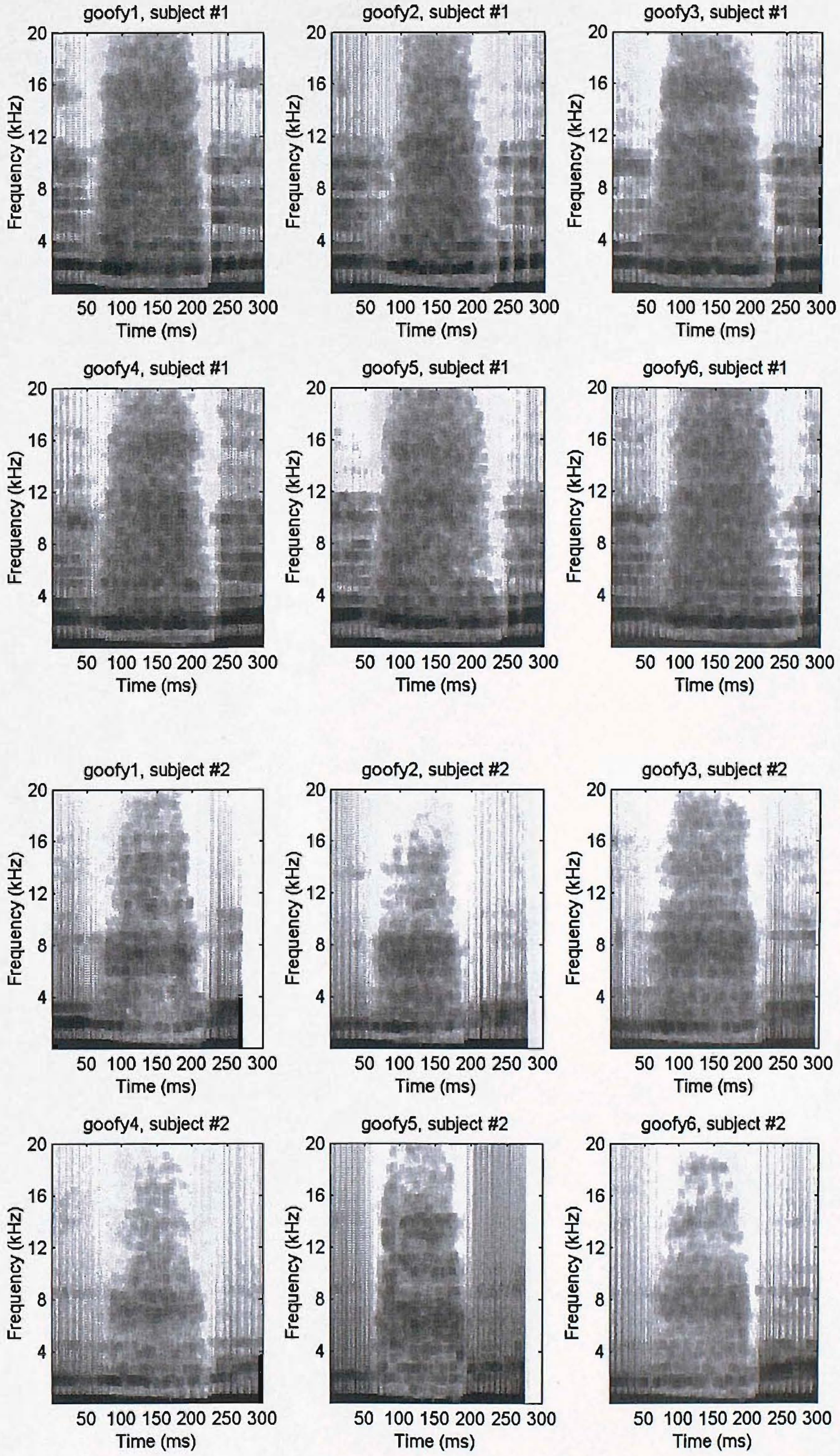
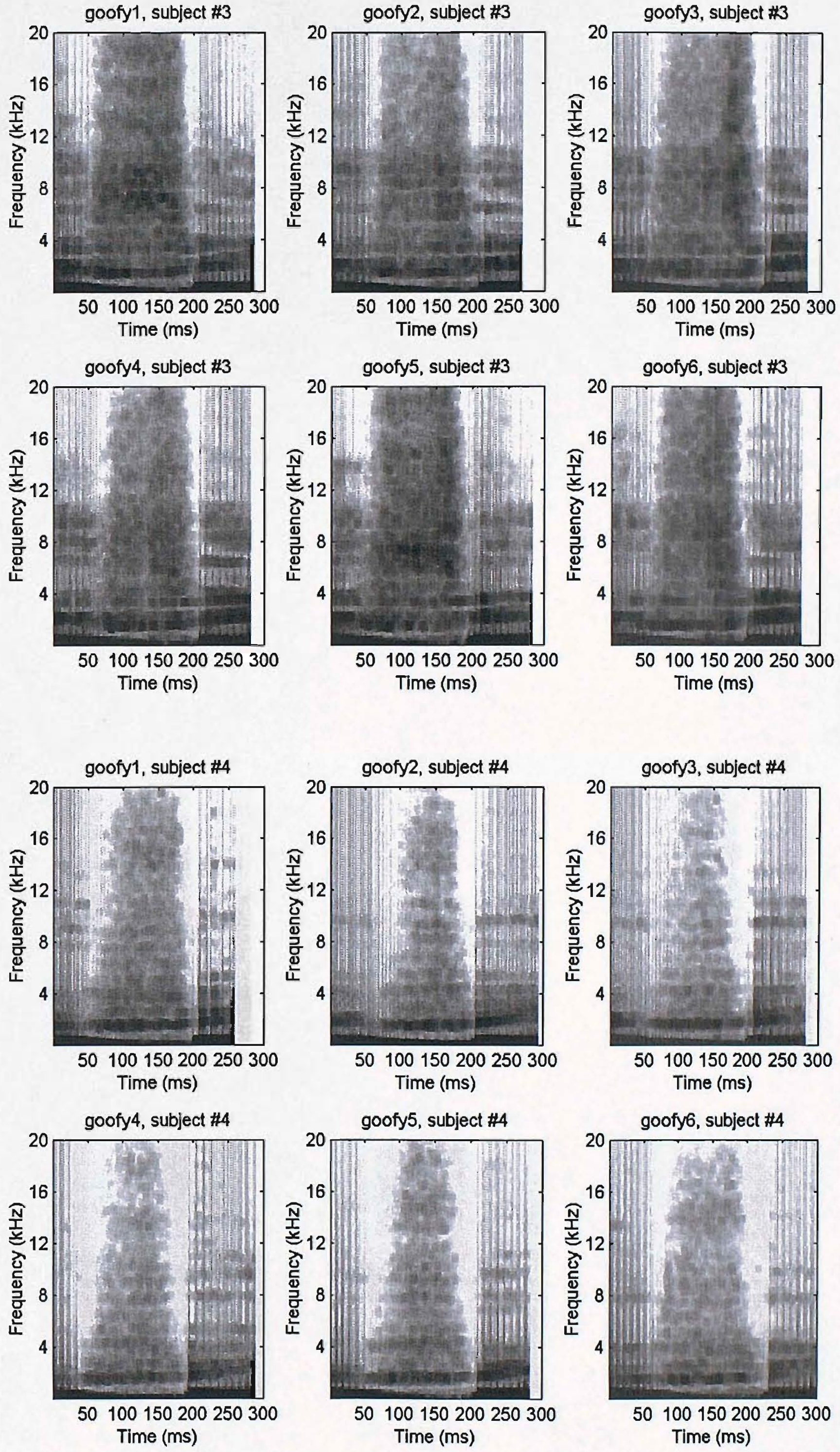


FIGURE B.22: Multitaper spectrograms: ['ufi] productions from "goofy", subjects M-01 and M-02.



goofy1, subject #3

goofy2, subject #3

goofy3, subject #3

goofy4, subject #3

goofy5, subject #3

goofy6, subject #3

goofy1, subject #4

goofy2, subject #4

goofy3, subject #4

goofy4, subject #4

goofy5, subject #4

goofy6, subject #4

FIGURE B.23: Multitaper spectrograms: ['ufi] productions from "goofy", subjects M-03 and M-04.

## B.2 Females

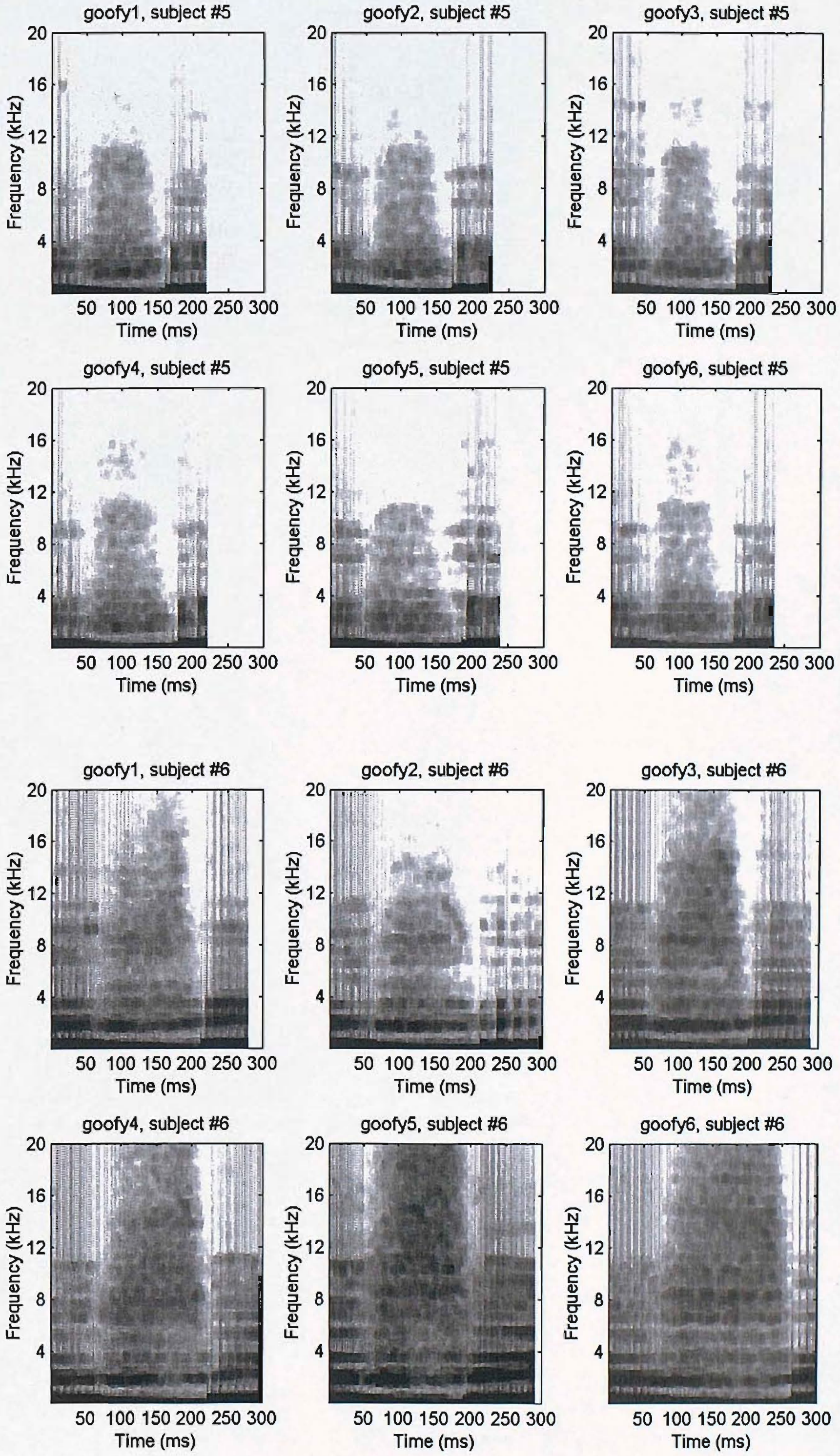


FIGURE B.24: Multitaper spectrograms: ['ufi] productions from "goofy", subjects M-05 and M-06.

## B.2 Females

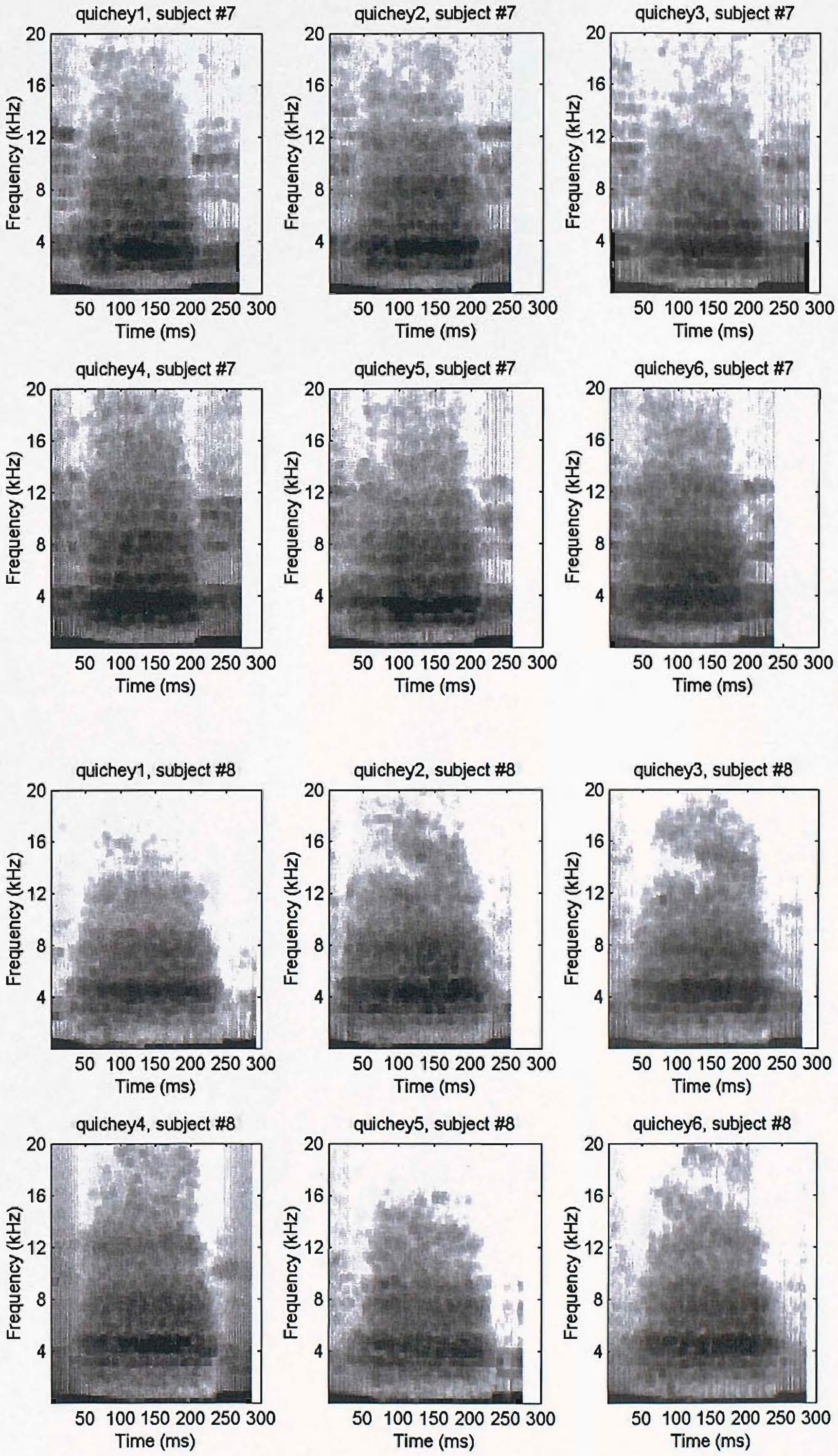


FIGURE B.25: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects F-07 and F-08.

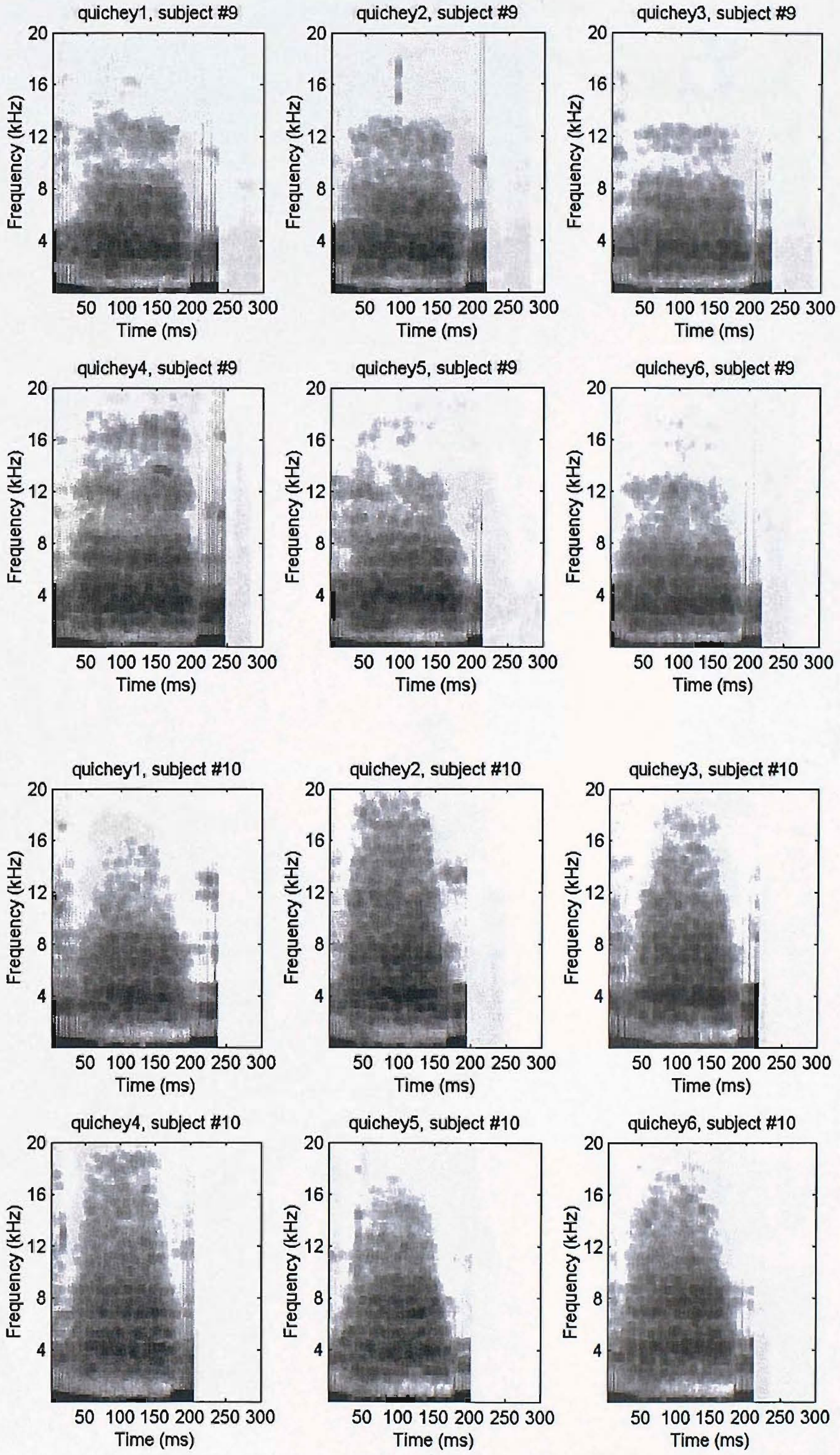


FIGURE B.26: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects F-09 and F-10.

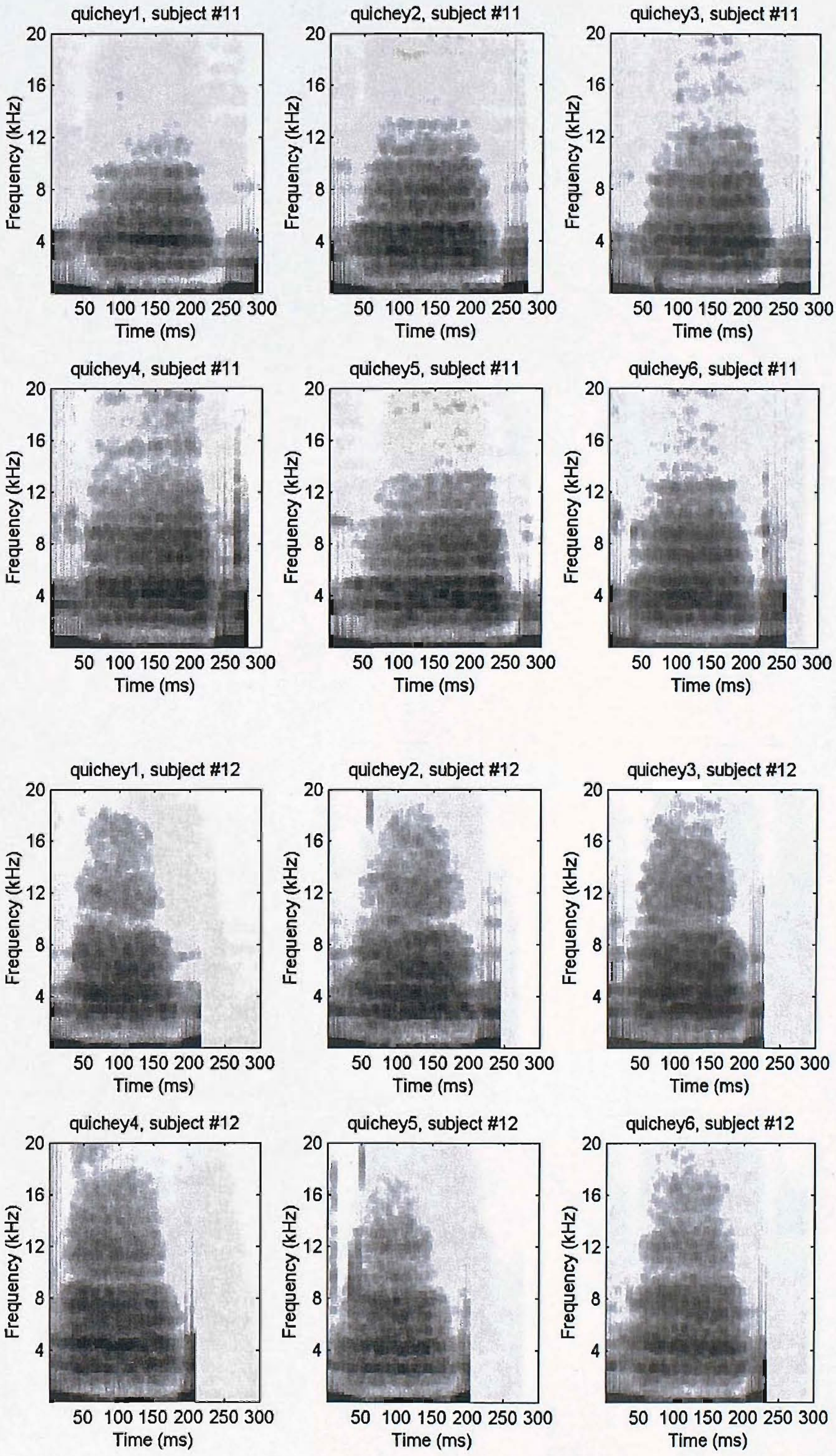


FIGURE B.27: Multitaper spectrograms: [ˈiʃi] productions from "quichey", subjects F-11 and F-12.

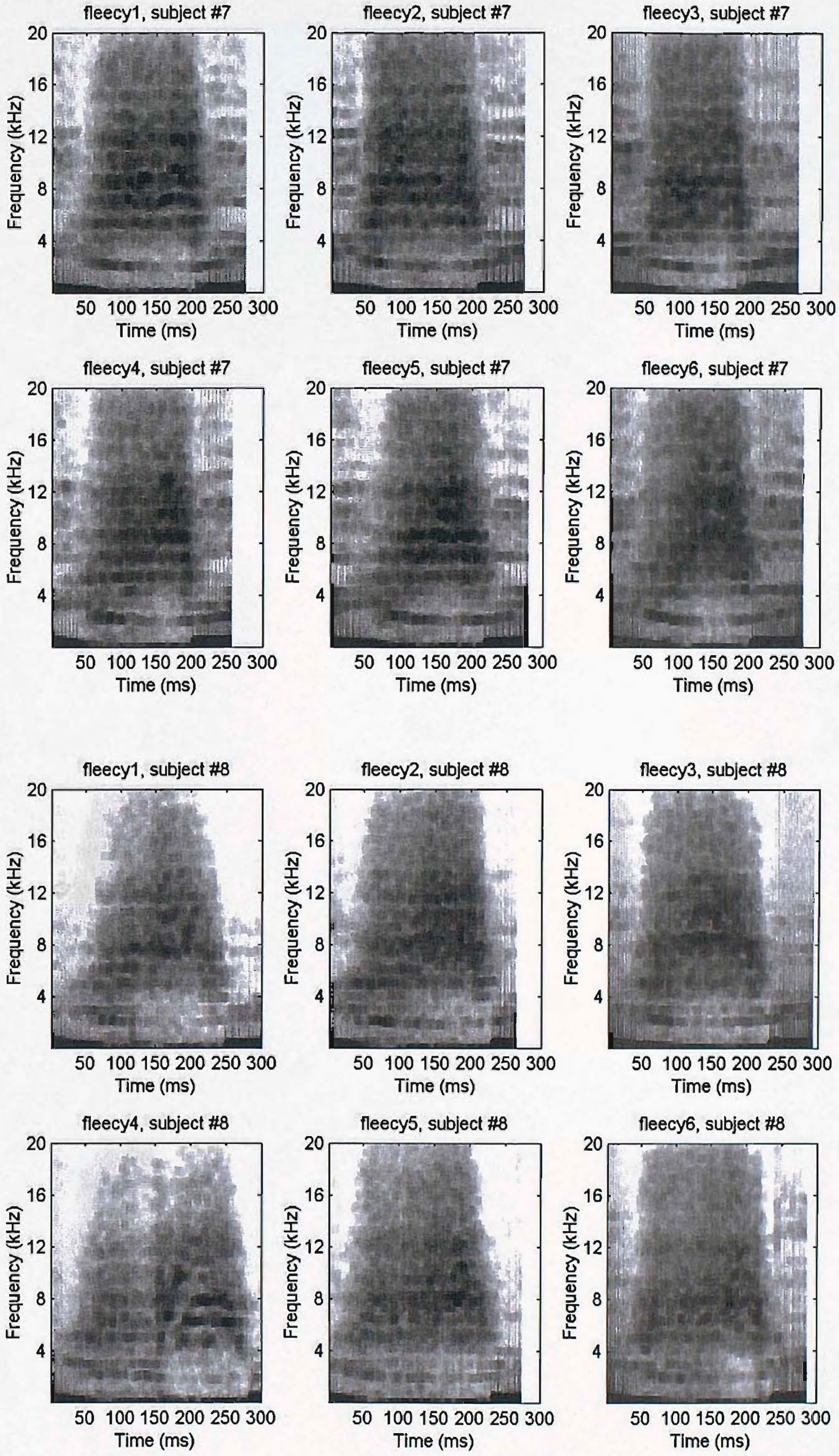


FIGURE B.28: Multitaper spectrograms: ['isi] productions from "fleecy", subjects F-07 and F-08.

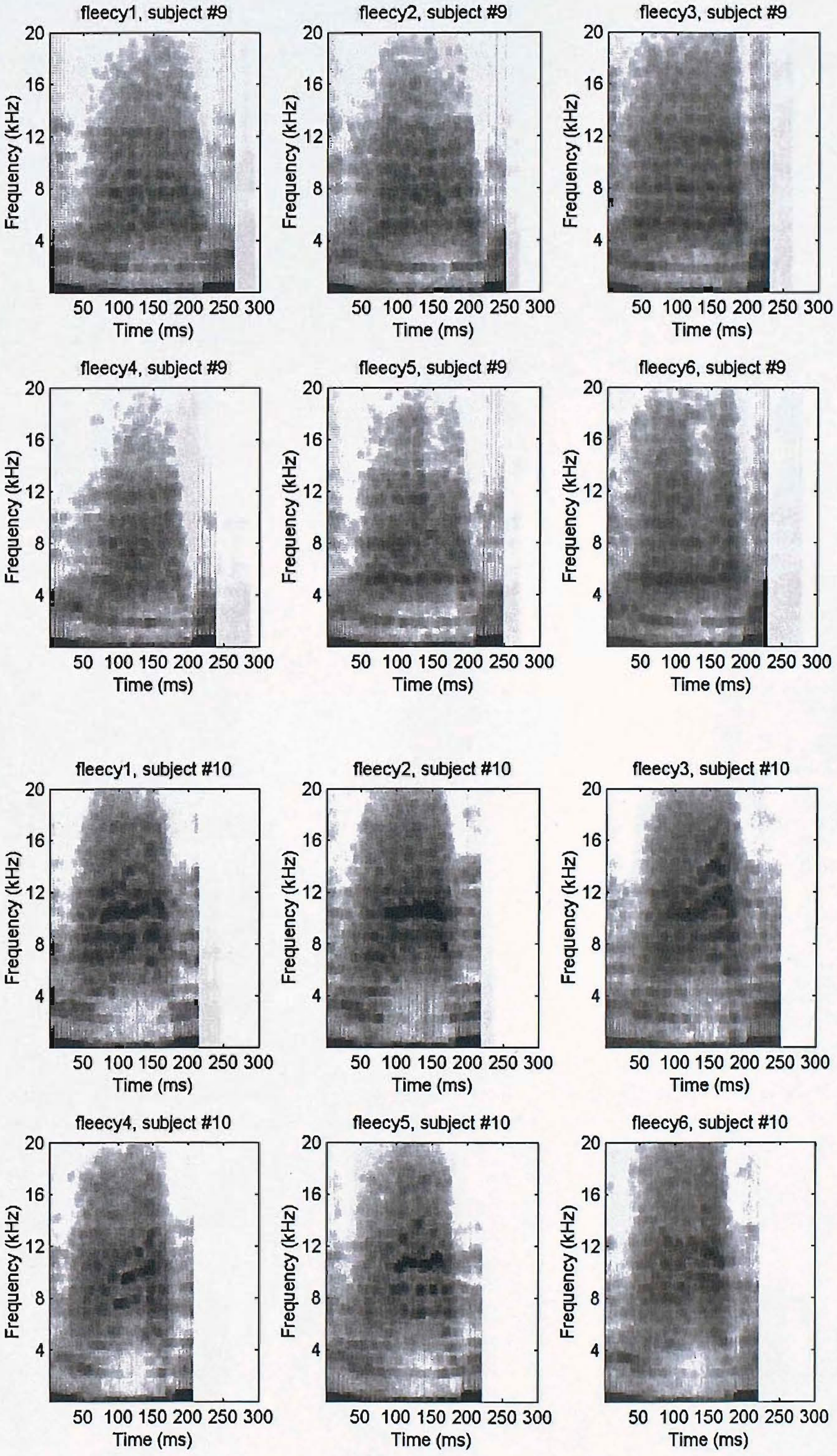


FIGURE B.29: Multitaper spectrograms: ['isi] productions from "fleecy", subjects F-09 and F-10.

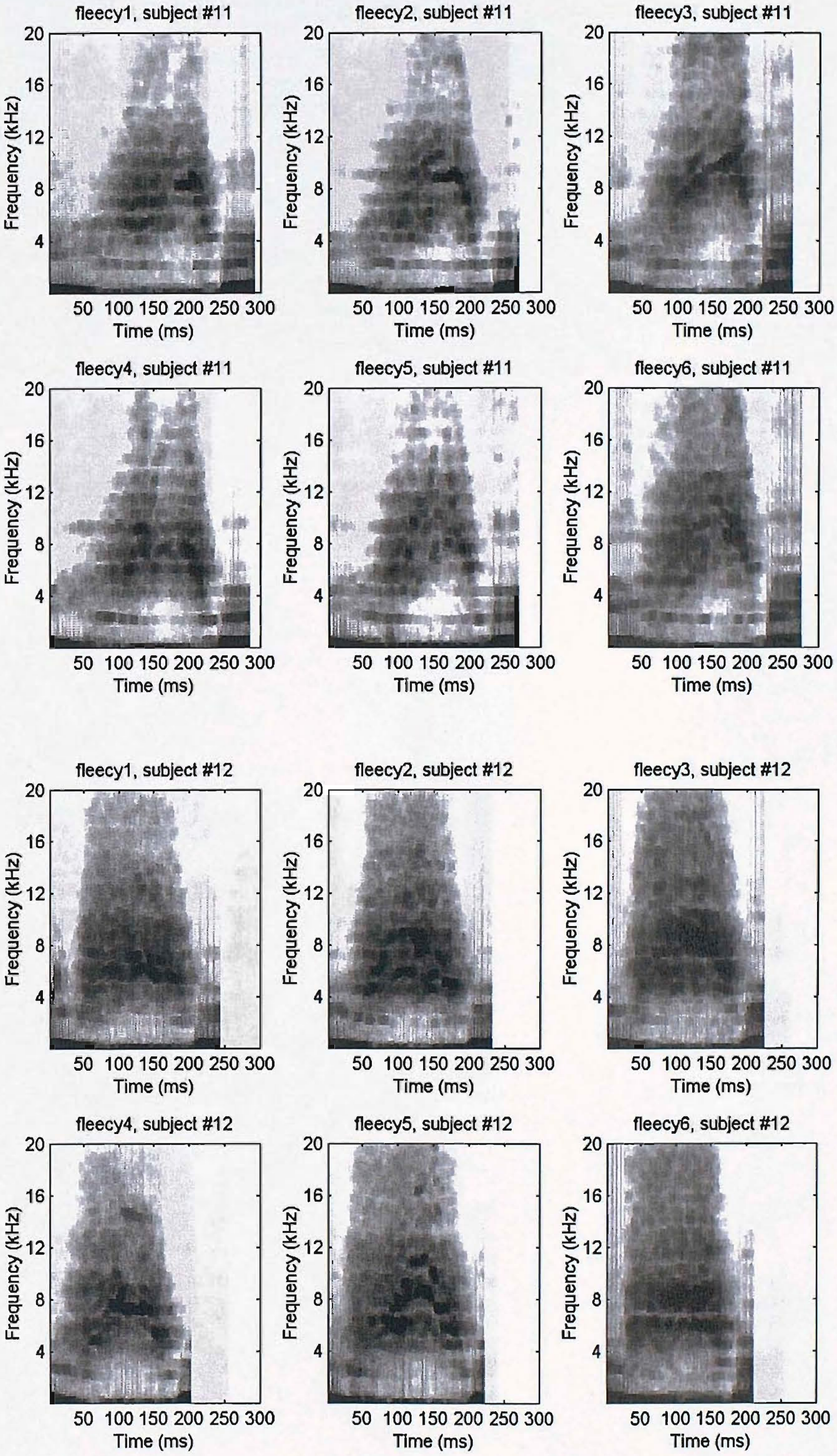


FIGURE B.30: Multitaper spectrograms: ['isi] productions from "fleecy", subjects F-11 and F-12.

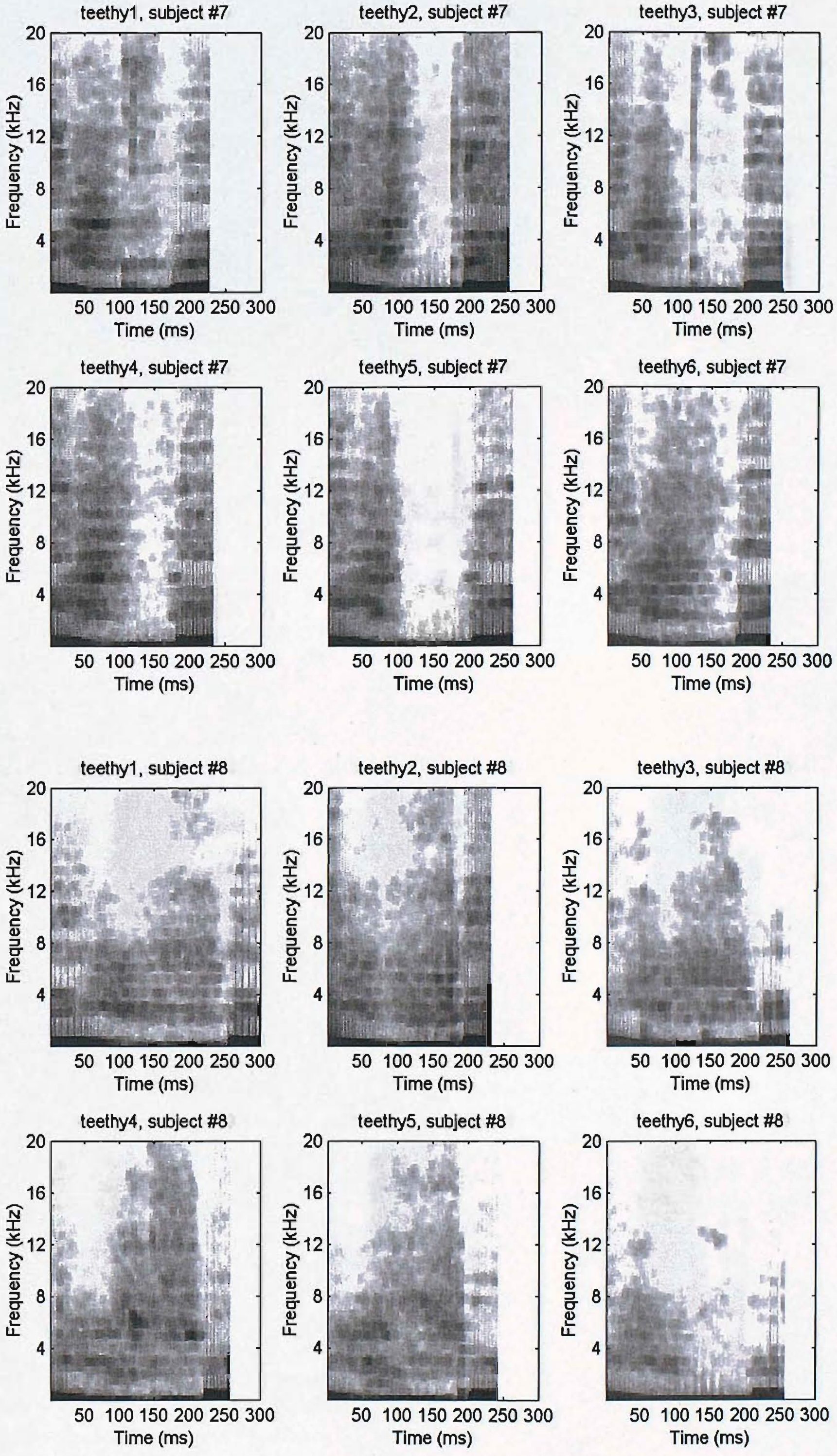


FIGURE B.31: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects F-07 and F-08.

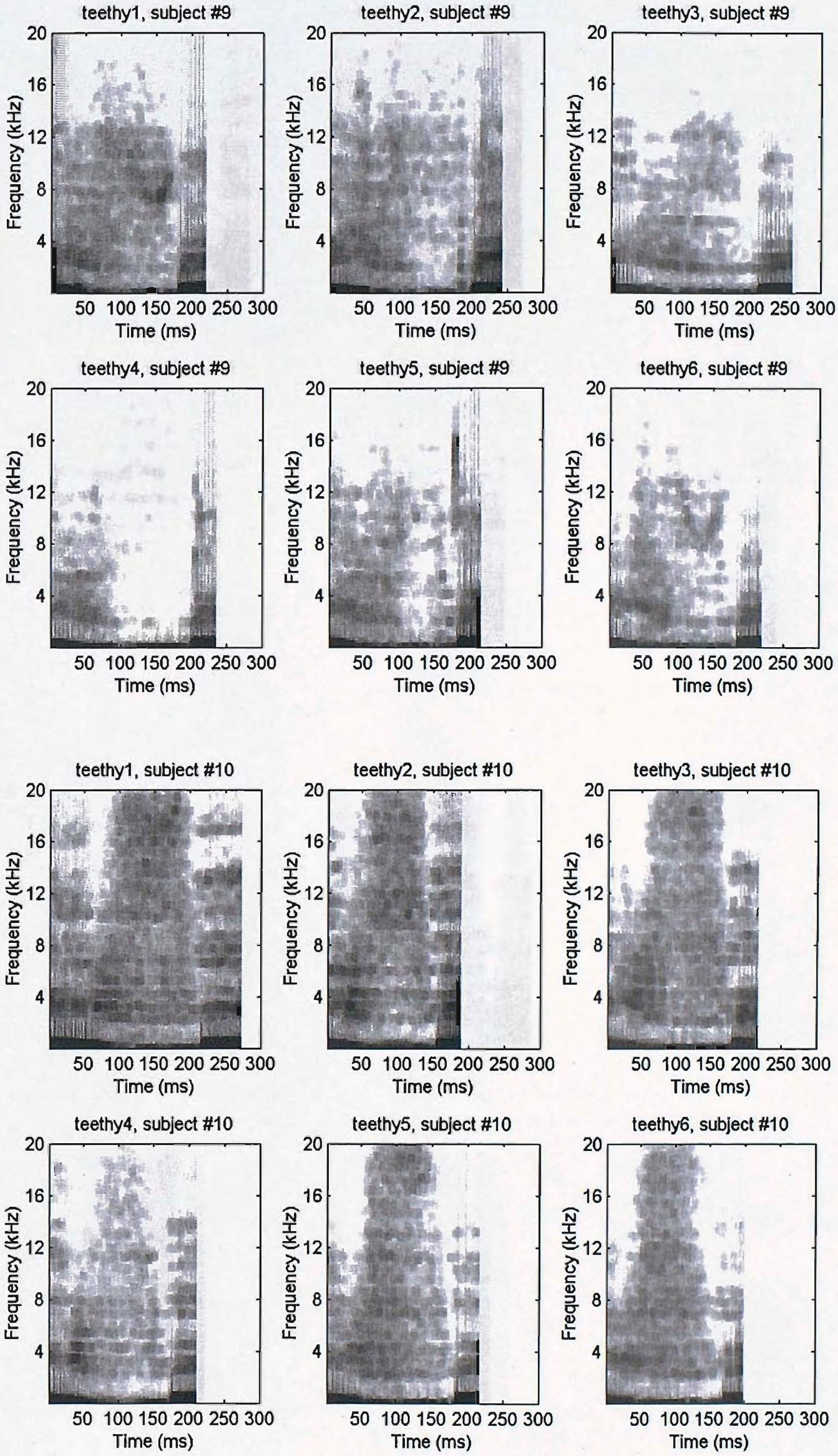


FIGURE B.32: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects F-09 and F-10.

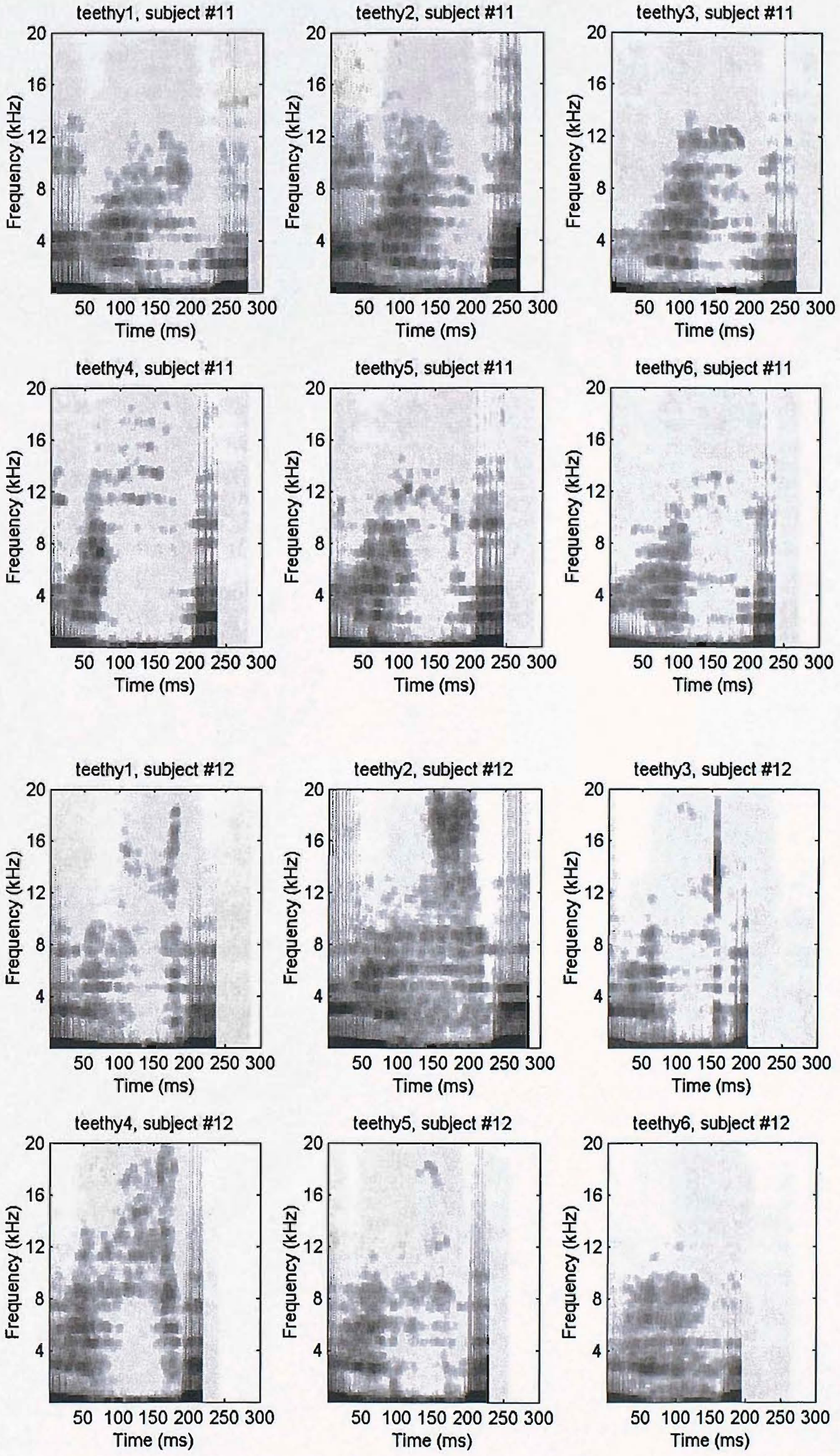


FIGURE B.33: Multitaper spectrograms: [ˈiθi] productions from "teethy", subjects F-11 and F-12.

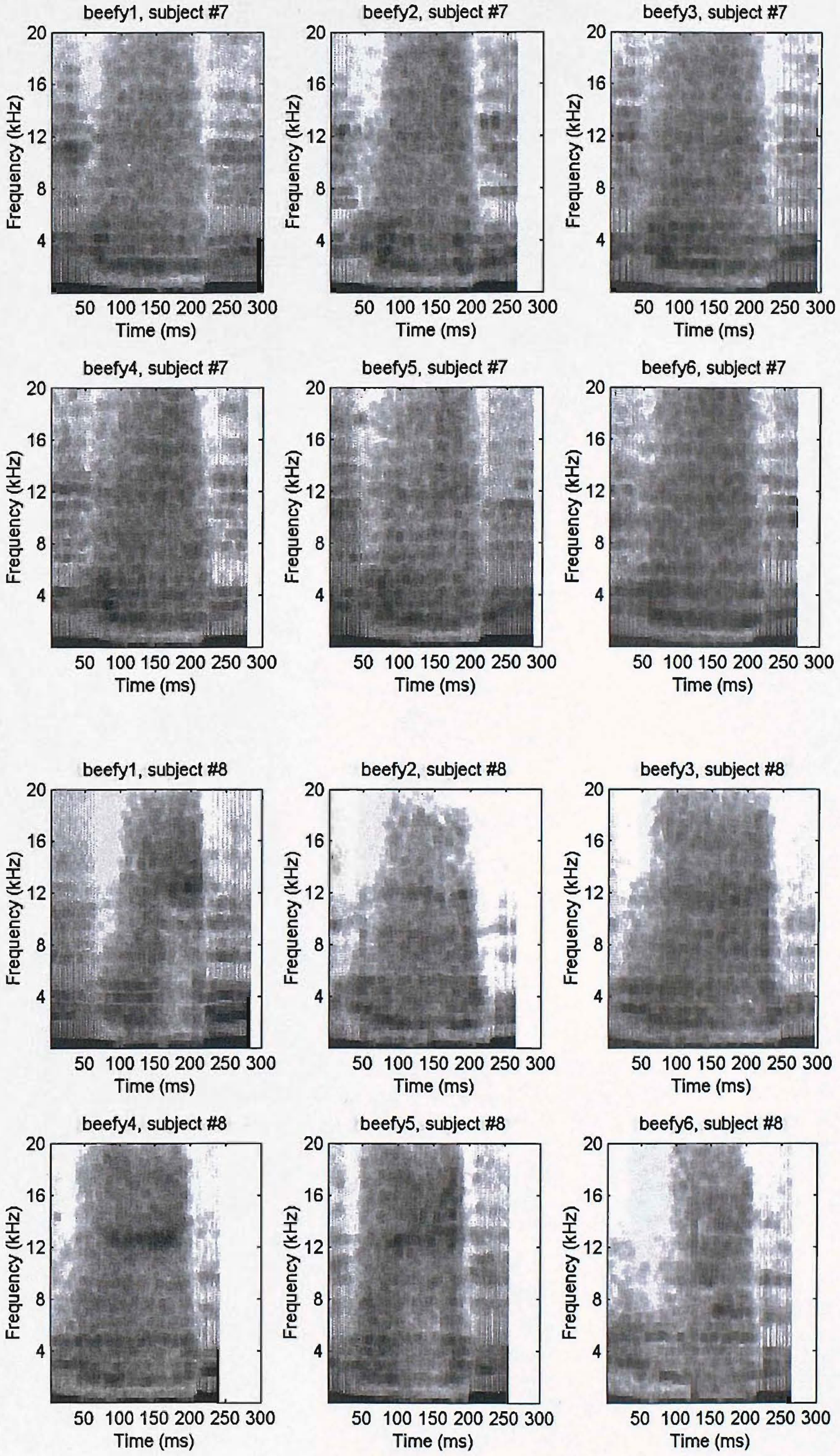


FIGURE B.34: Multitaper spectrograms: ['ifi] productions from "beefy", subjects F-07 and F-08.

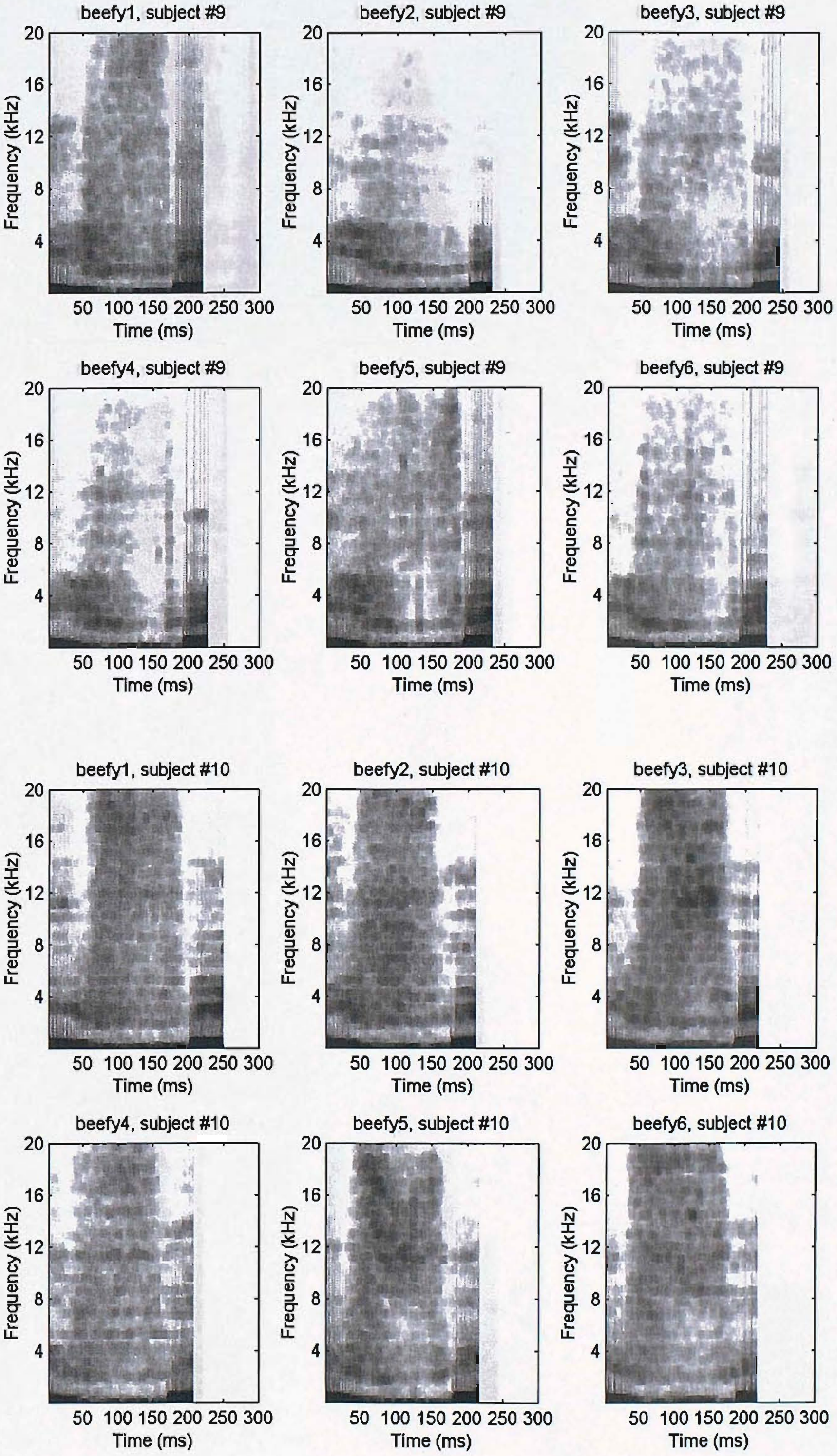


FIGURE B.35: Multitaper spectrograms: ['ifi] productions from "beefy", subjects F-09 and F-10.

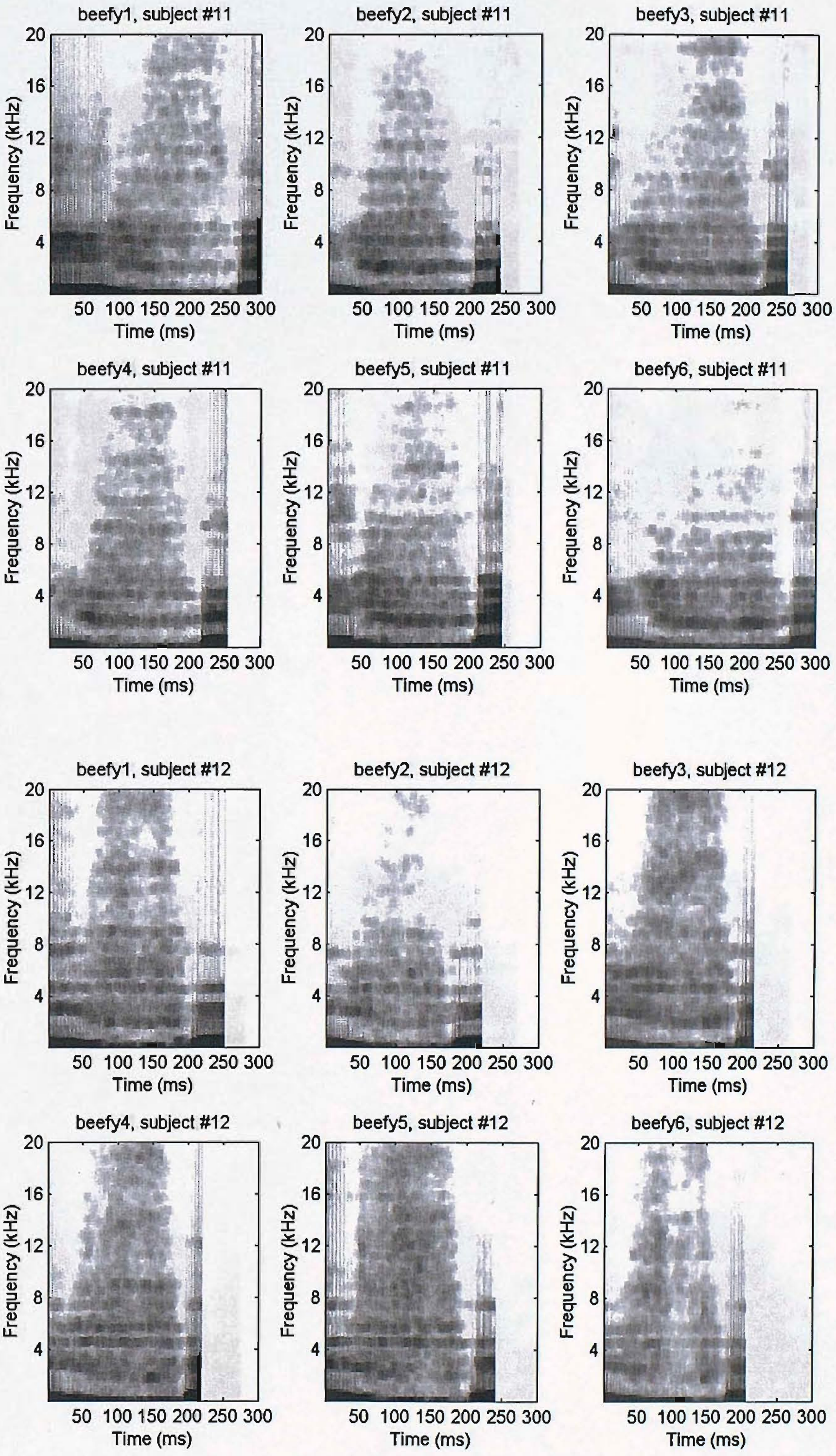


FIGURE B.36: Multitaper spectrograms: ['ifi] productions from "beefy", subjects F-11 and F-12.

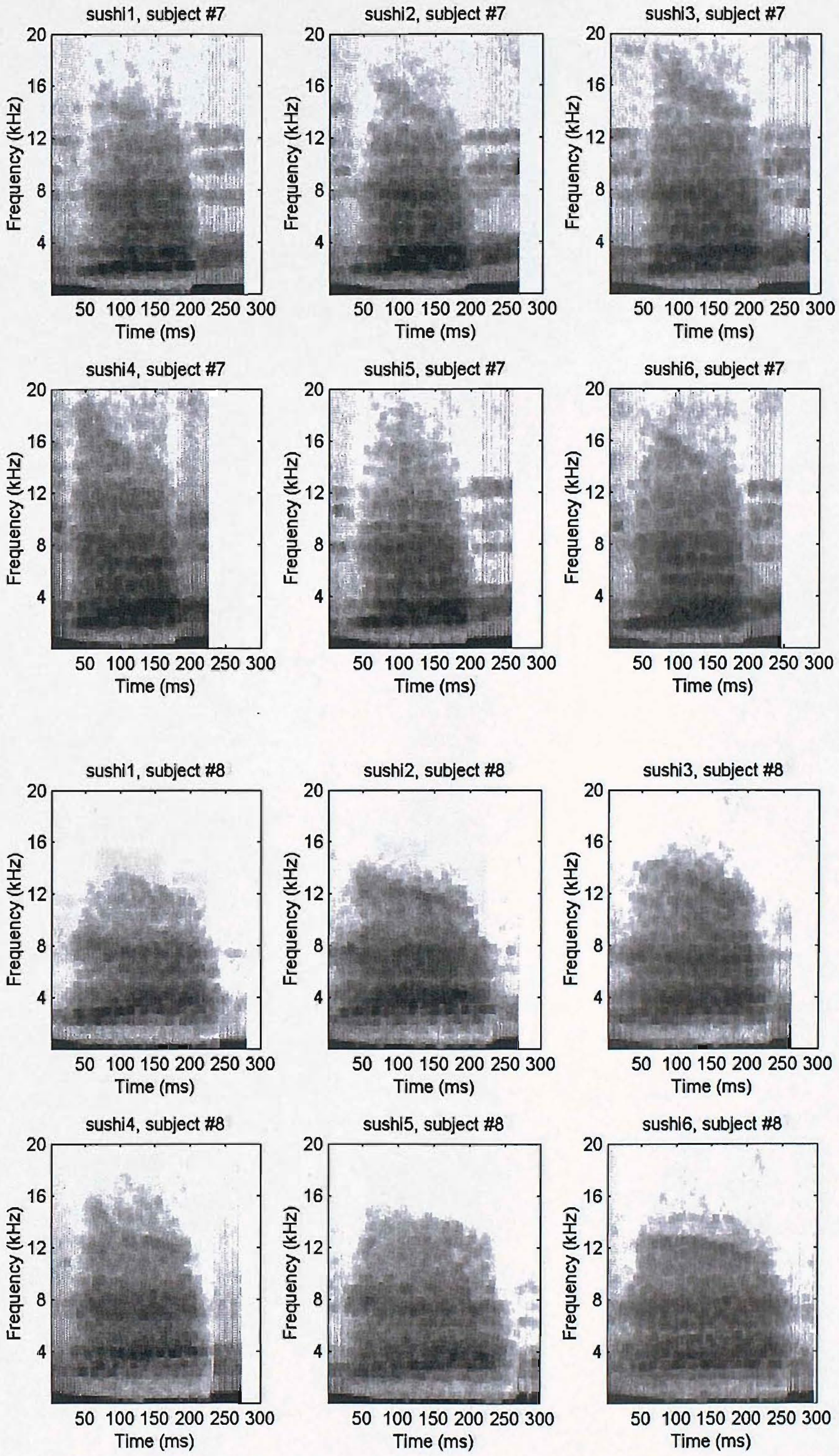


FIGURE B.37: Multitaper spectrograms: ['uʃi] productions from “sushi, subjects F-07 and F-08.

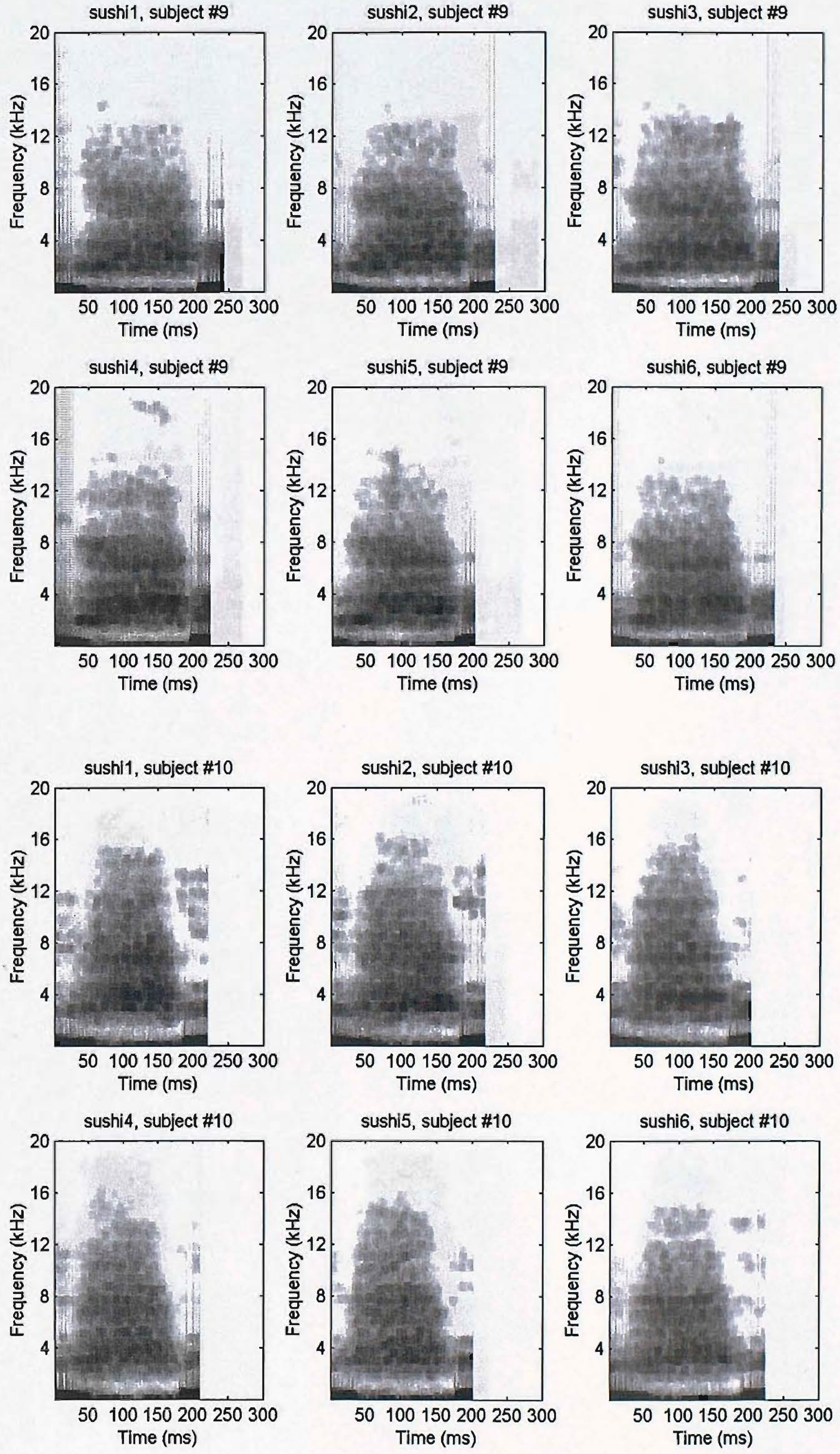


FIGURE B.38: Multitaper spectrograms: ['uʃi] productions from “sushi, subjects F-09 and F-10.

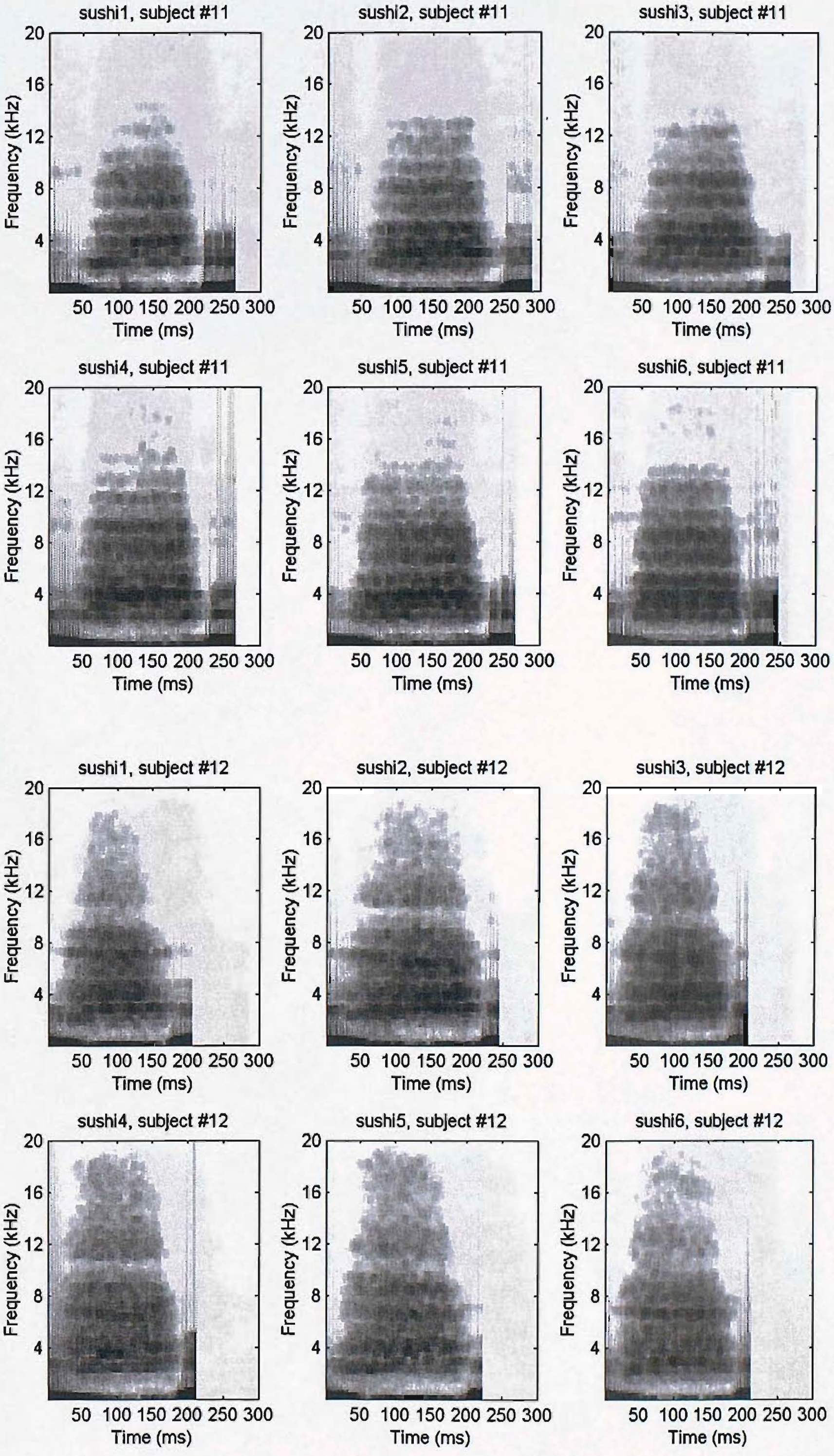


FIGURE B.39: Multitaper spectrograms: ['uʃi] productions from "sushi, subjects F-11 and F-12.

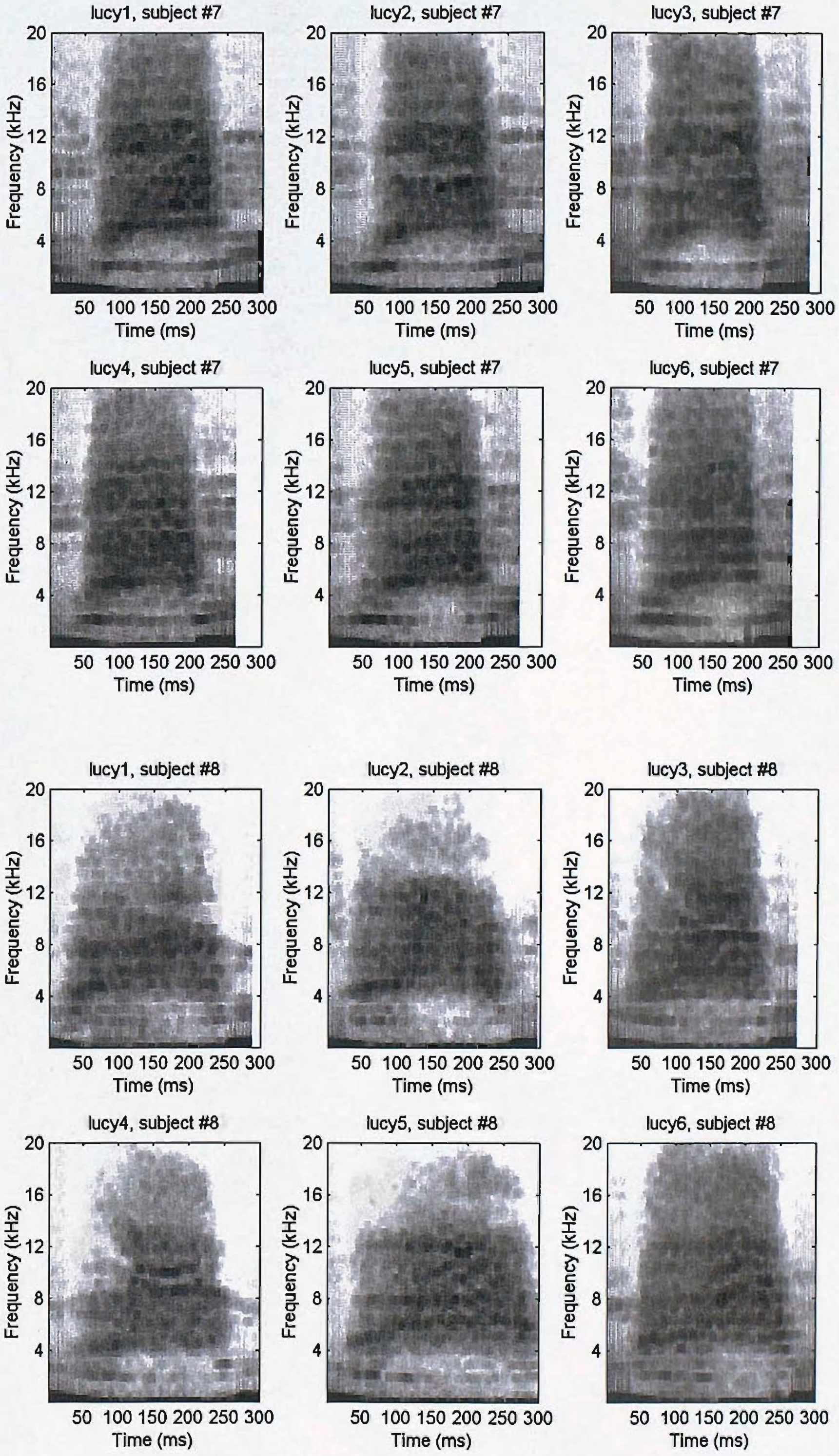
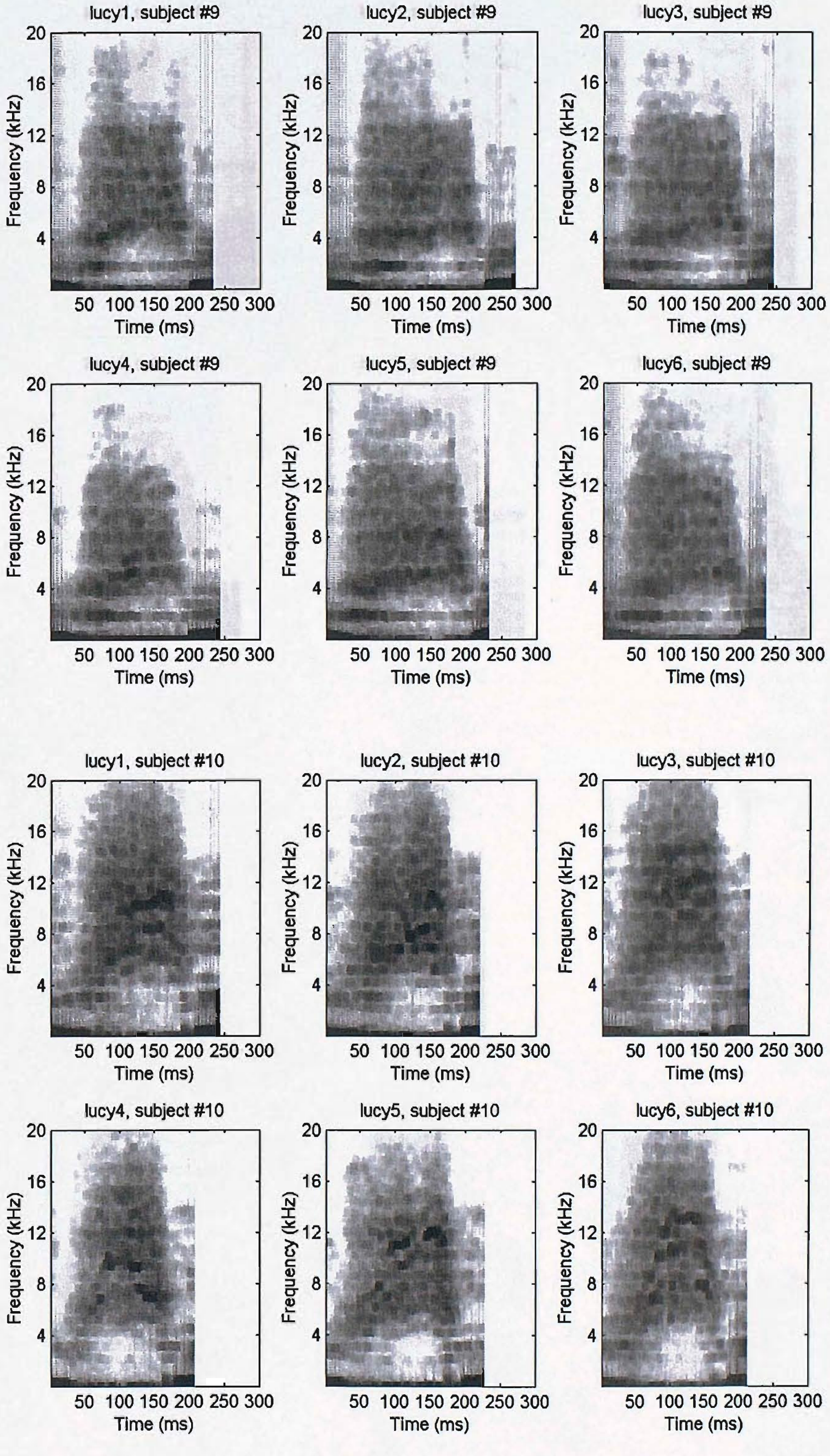


FIGURE B.40: Multitaper spectrograms: ['usi] productions from "Lucy", subjects F-07 and F-08.



lucy1, subject #9

lucy2, subject #9

lucy3, subject #9

lucy4, subject #9

lucy5, subject #9

lucy6, subject #9

lucy1, subject #10

lucy2, subject #10

lucy3, subject #10

lucy4, subject #10

lucy5, subject #10

lucy6, subject #10

FIGURE B.41: Multitaper spectrograms: ['usi] productions from "Lucy", subjects F-09 and F-10.

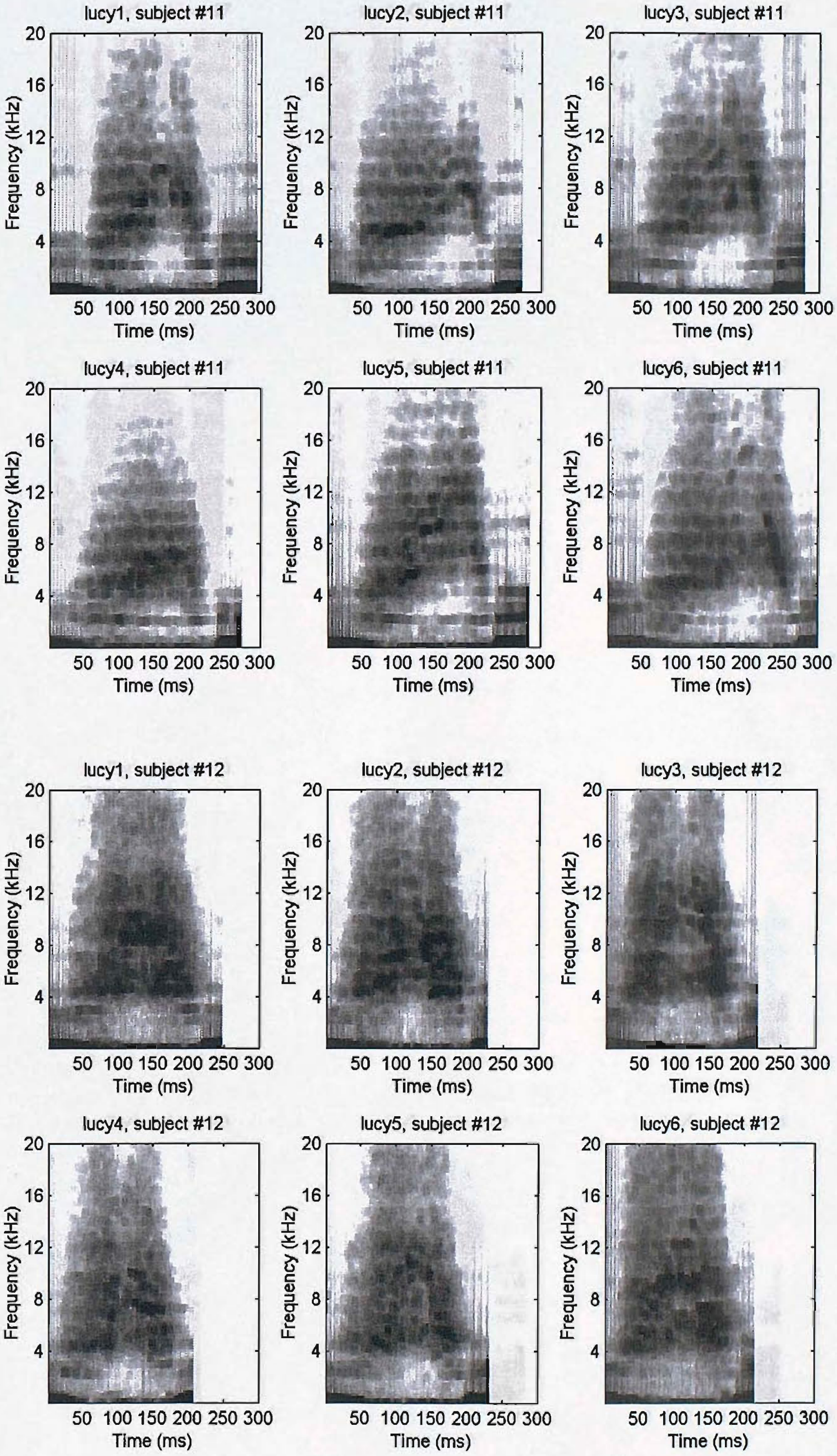


FIGURE B.42: Multitaper spectrograms: ['usi] productions from "Lucy", subjects F-11 and F-12.

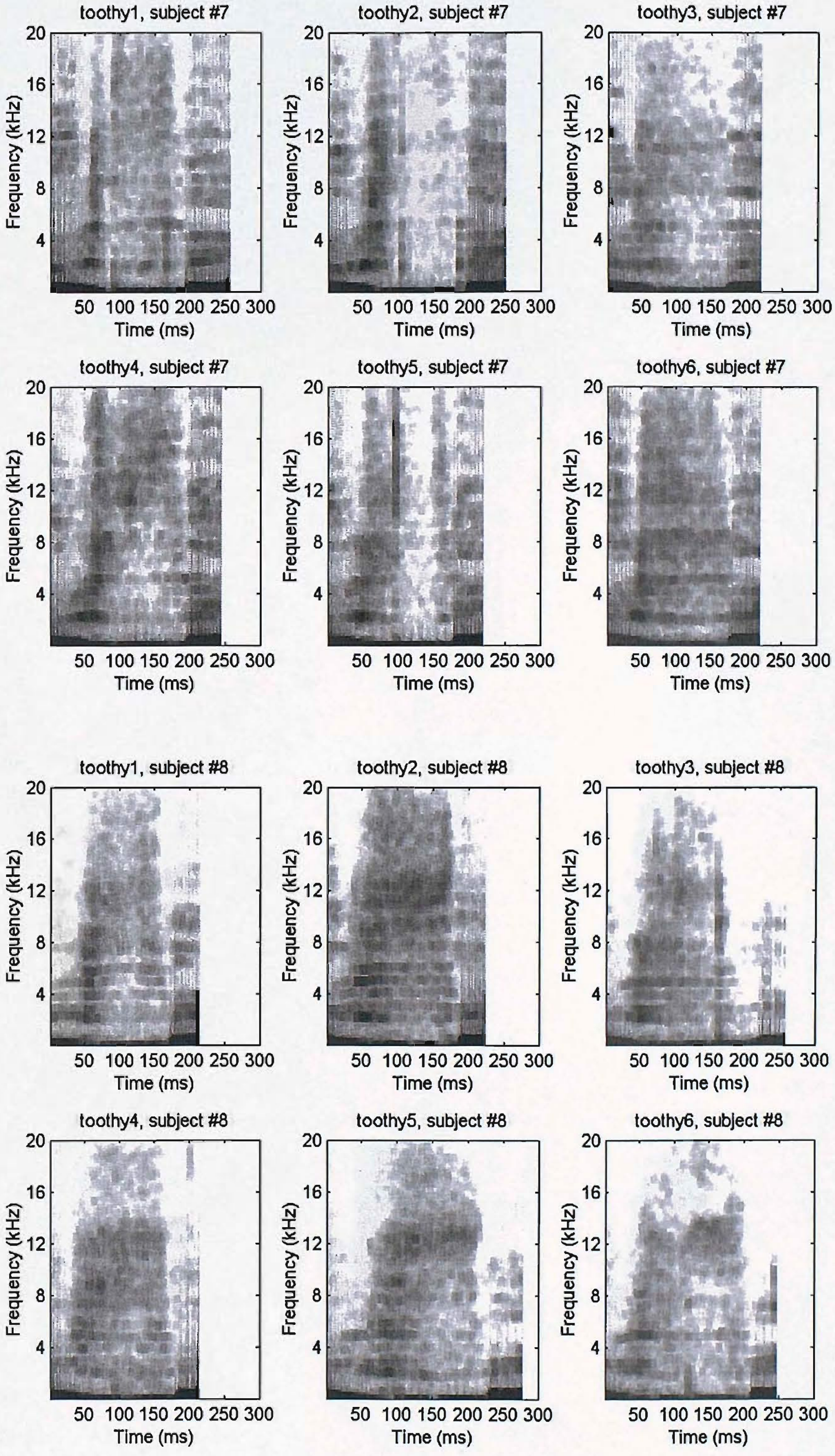


FIGURE B.43: Multitaper spectrograms: [ˈuθi] productions from "toothy", subjects F-07 and F-08.

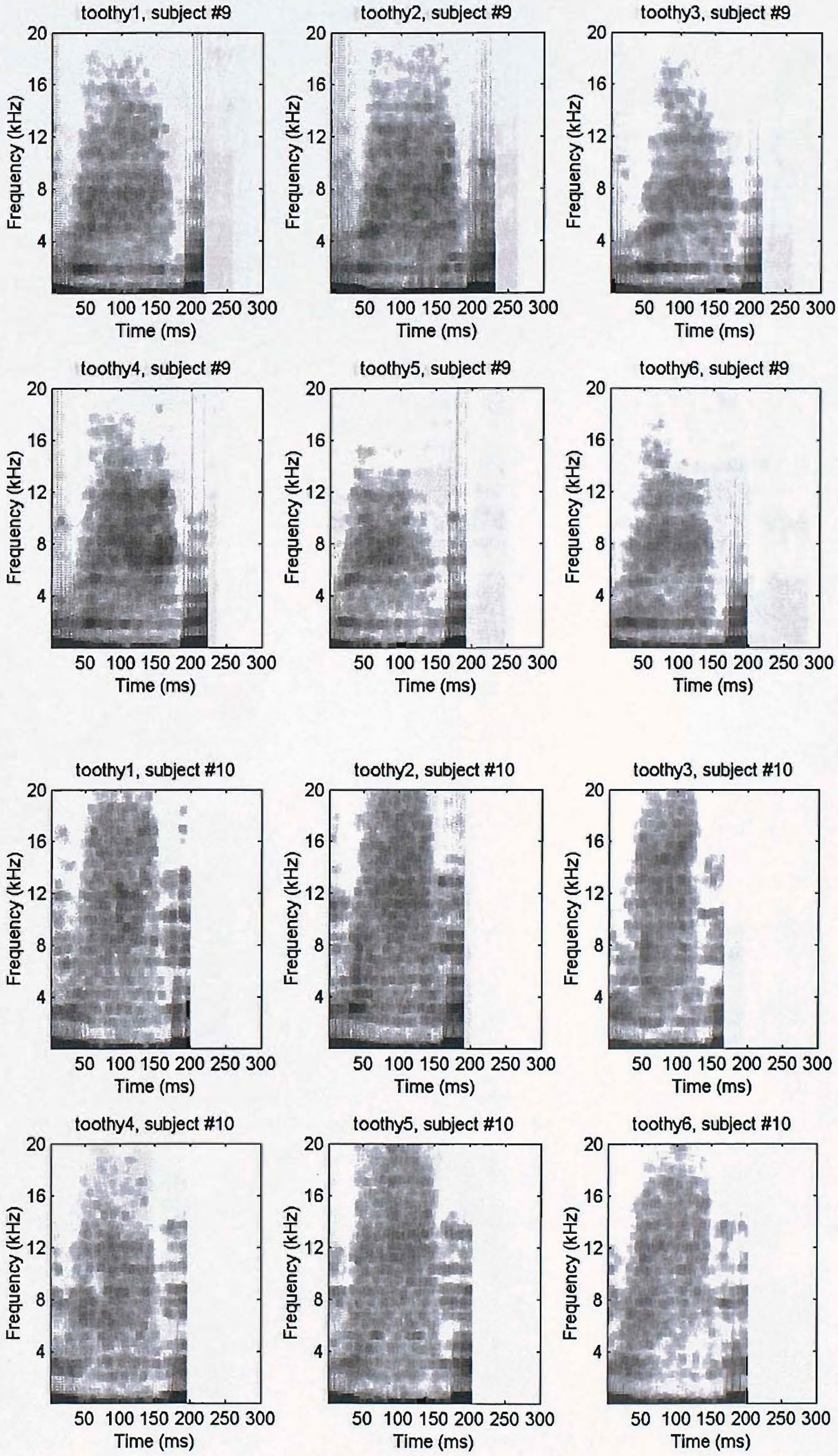


FIGURE B.44: Multitaper spectrograms: ['uθi] productions from "toothy", subjects F-09 and F-10.

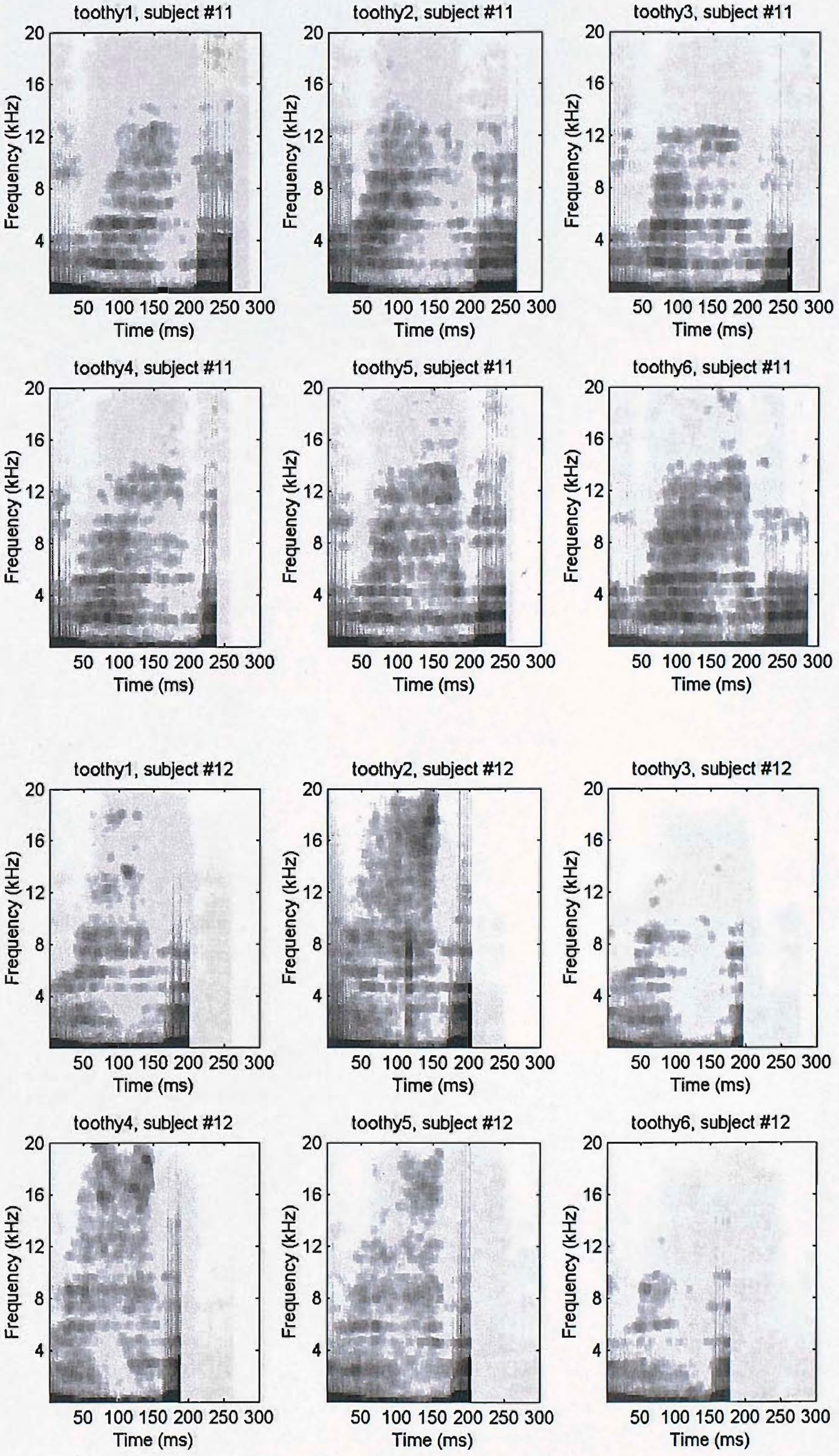
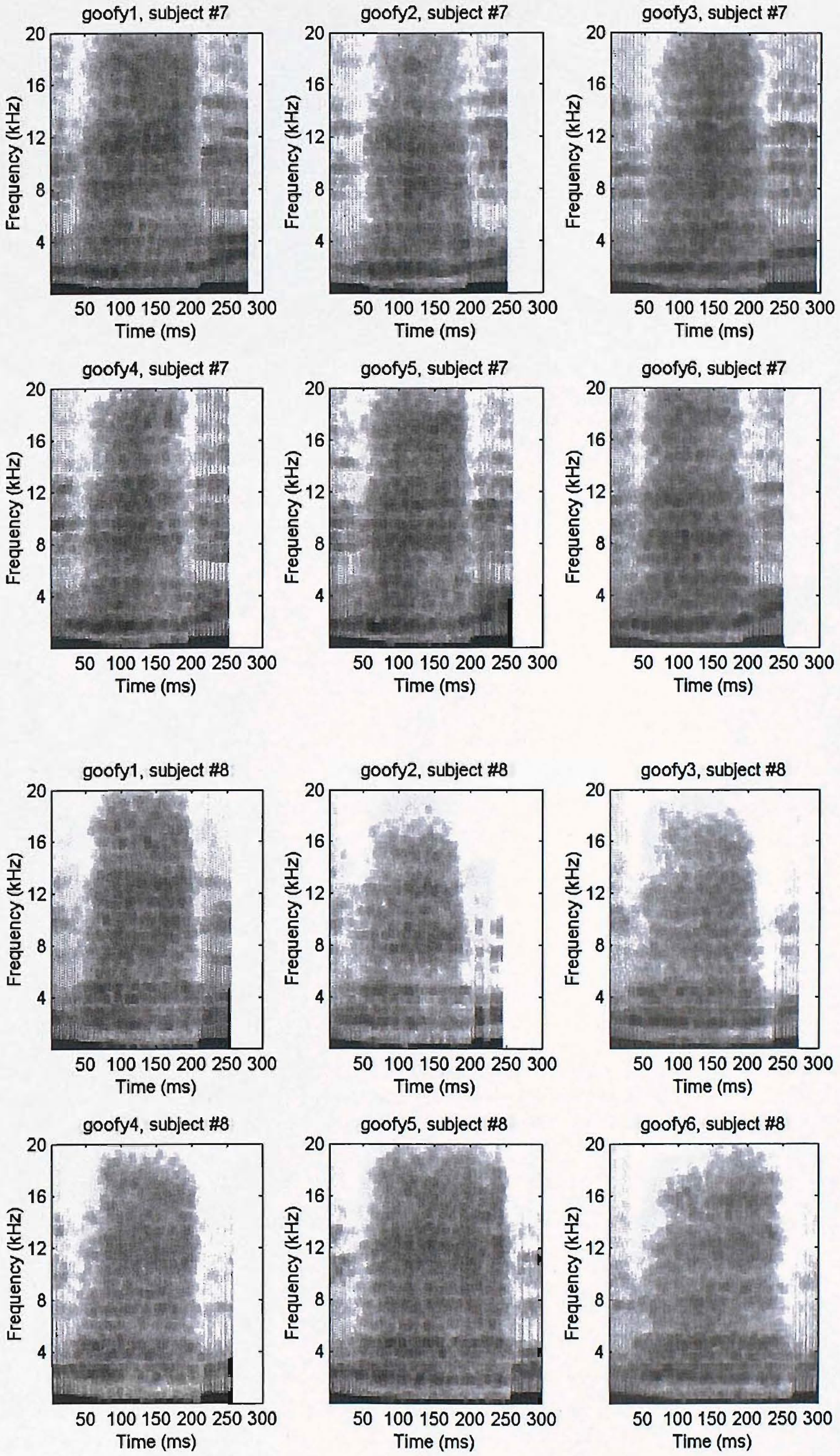


FIGURE B.45: Multitaper spectrograms: ['uθi] productions from "toothy", subjects F-11 and F-12.



goofy1, subject #7

goofy2, subject #7

goofy3, subject #7

goofy4, subject #7

goofy5, subject #7

goofy6, subject #7

goofy1, subject #8

goofy2, subject #8

goofy3, subject #8

goofy4, subject #8

goofy5, subject #8

goofy6, subject #8

FIGURE B.46: Multitaper spectrograms: ['ufi] productions from "goofy", subjects F-07 and F-08.

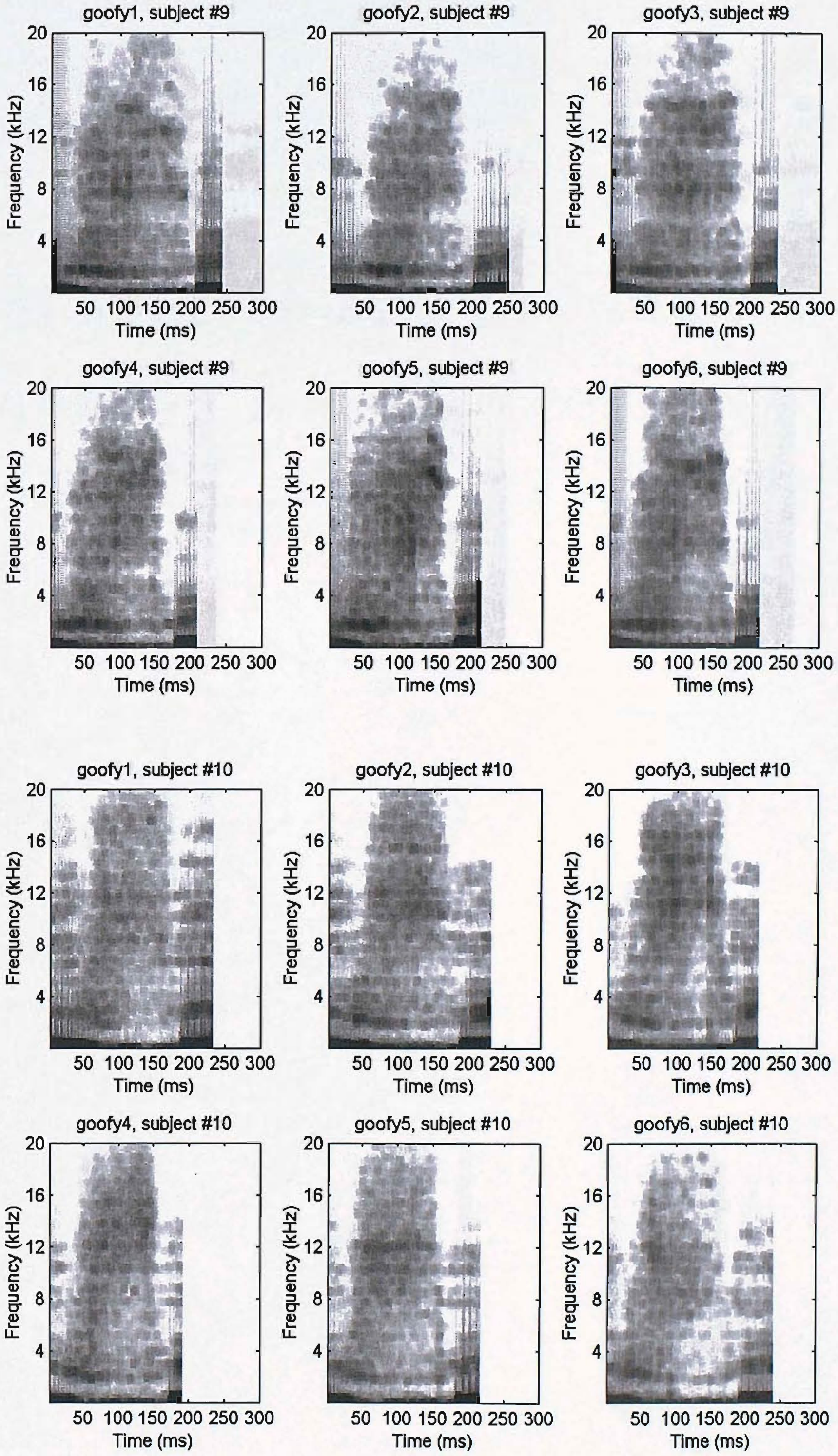


FIGURE B.47: Multitaper spectrograms: ['ufi] productions from "goofy", subjects F-09 and F-10.

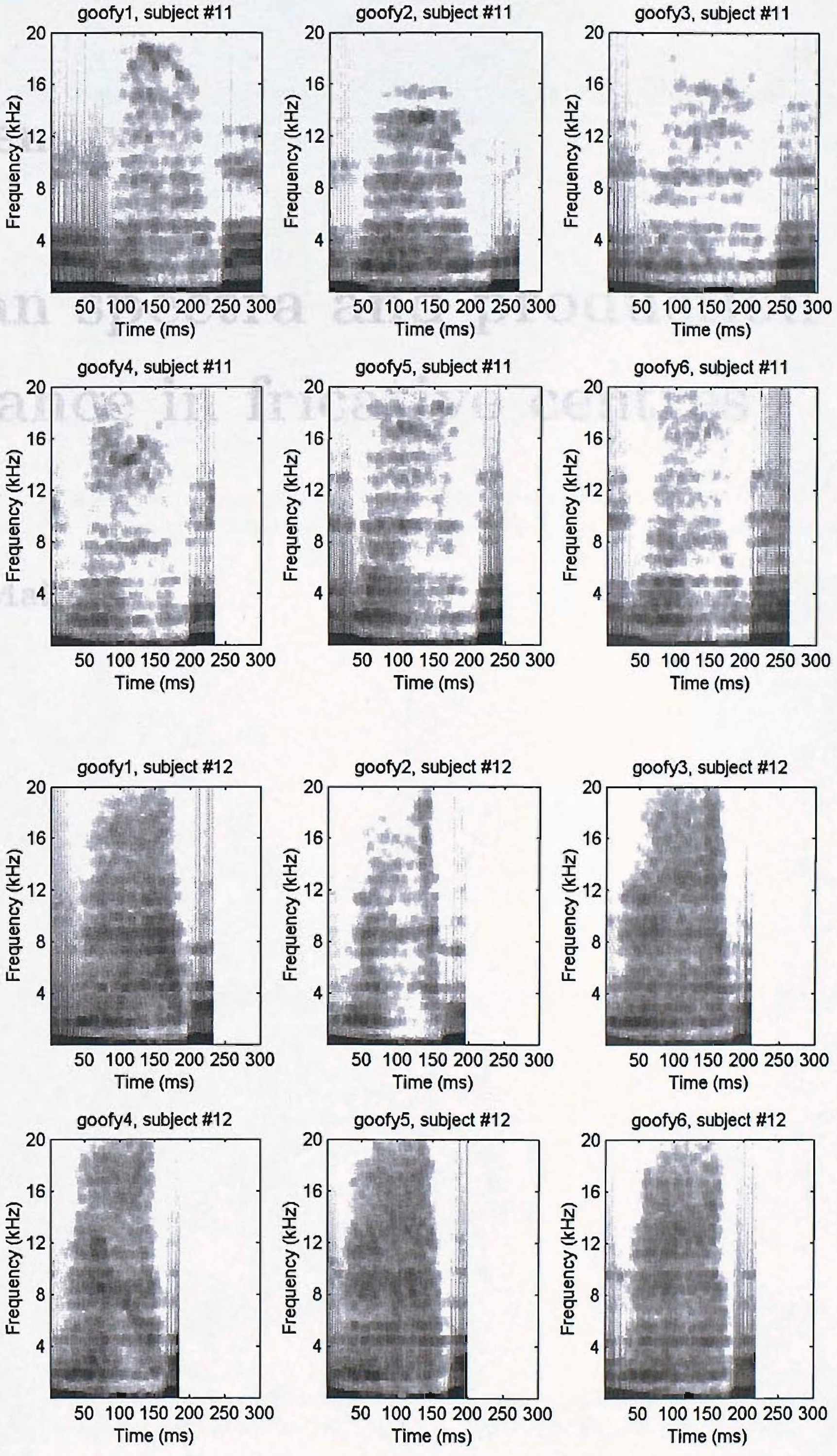


FIGURE B.48: Multitaper spectrograms: [ˈufi] productions from "goofy", subjects F-11 and F-12.

# Appendix C

# Mean spectra and production variance in fricative centres

## C.1 Males

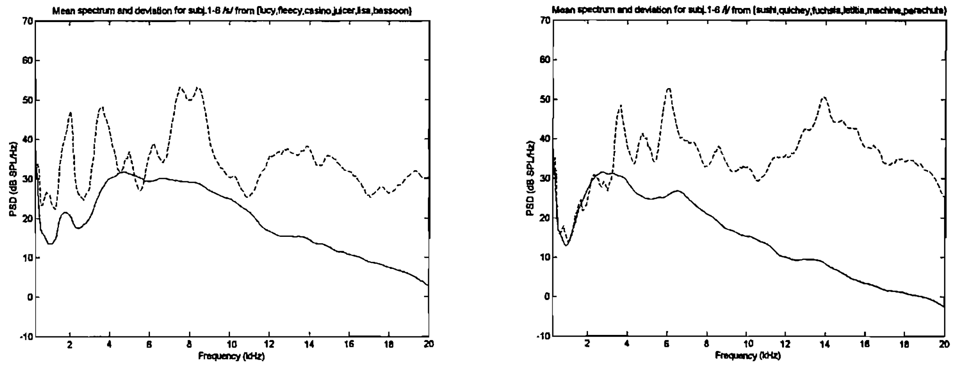


FIGURE C.1: Mean spectrum (solid) and production variance (dashed) from central /s/ (left) in /usi,isi,əsi,isə,əsu,usə/ contexts, and /ʃ/ (right) in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, all male subjects.

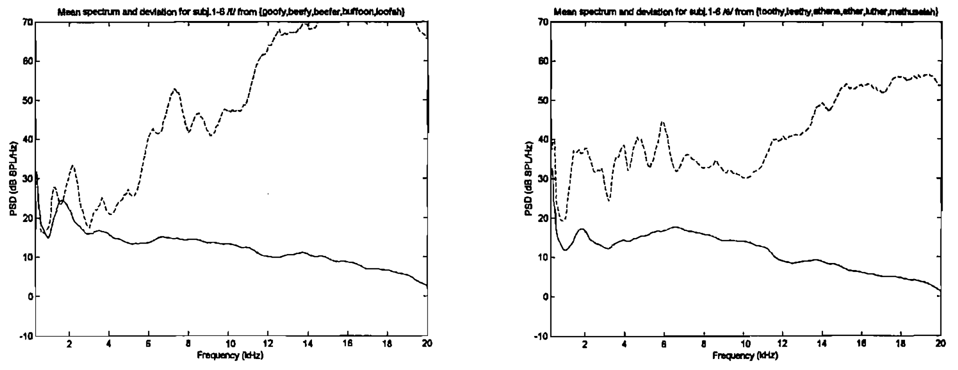


FIGURE C.2: Mean spectrum (solid) and production variance (dashed) from central /f/ (left) in /ufi,ifi,ifə,əfu,ufə/ contexts, and /θ/ (right) in /uθi,iθi,əθi,iθə,uθə,əθu/ contexts, all male subjects.

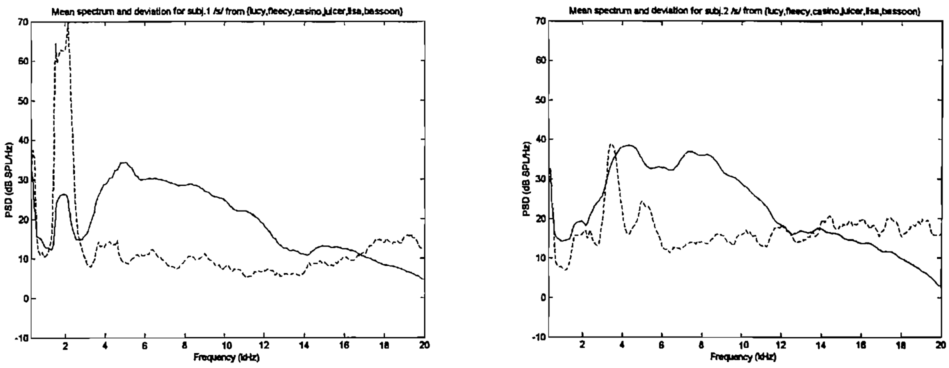


FIGURE C.3: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject M-01 (left) and M-02 (right).

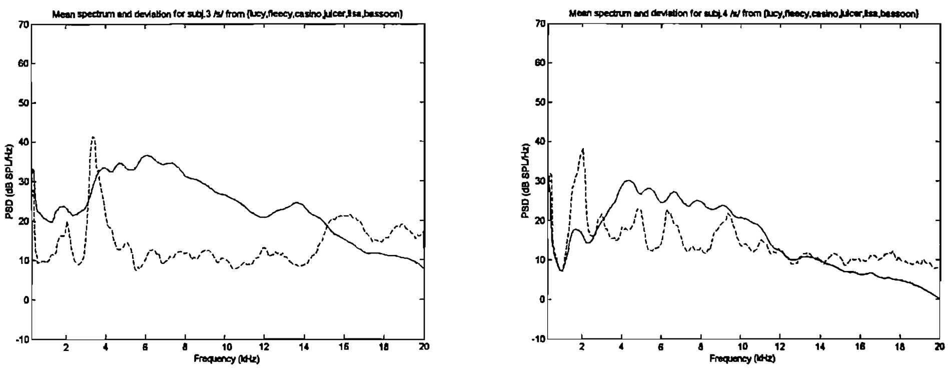


FIGURE C.4: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject M-03 (left) and M-04 (right).

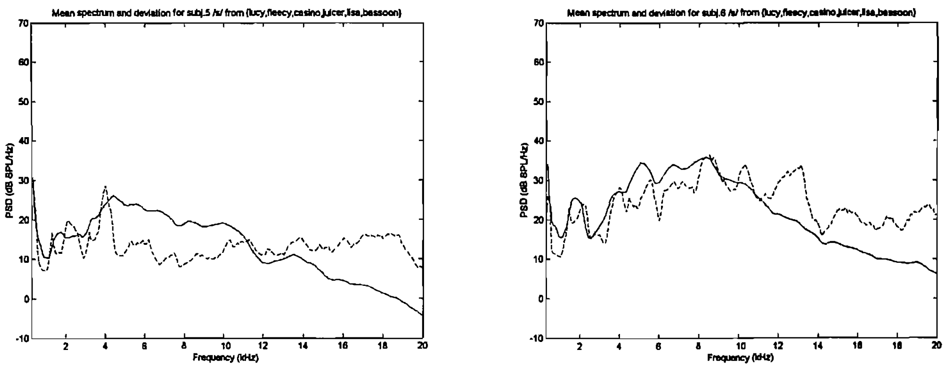


FIGURE C.5: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject M-05 (left) and M-06 (right).

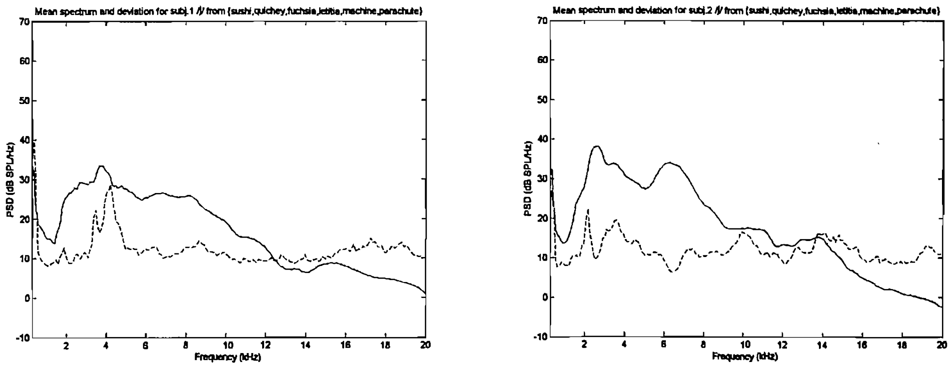


FIGURE C.6: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject M-01 (left) and M-02 (right).

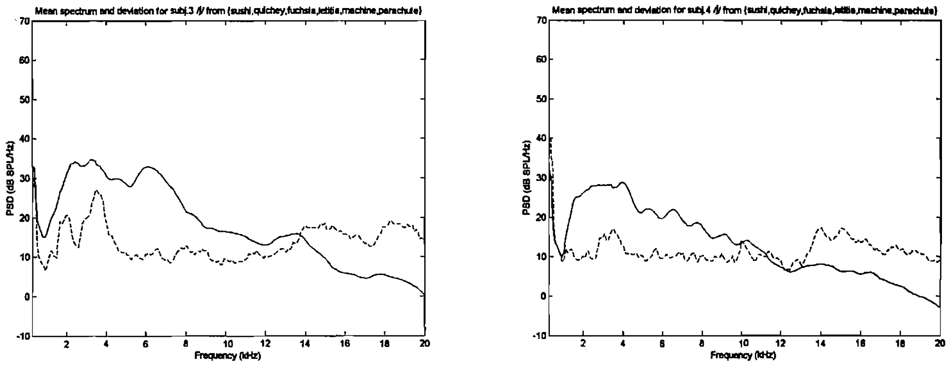


FIGURE C.7: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject M-03 (left) and M-04 (right).

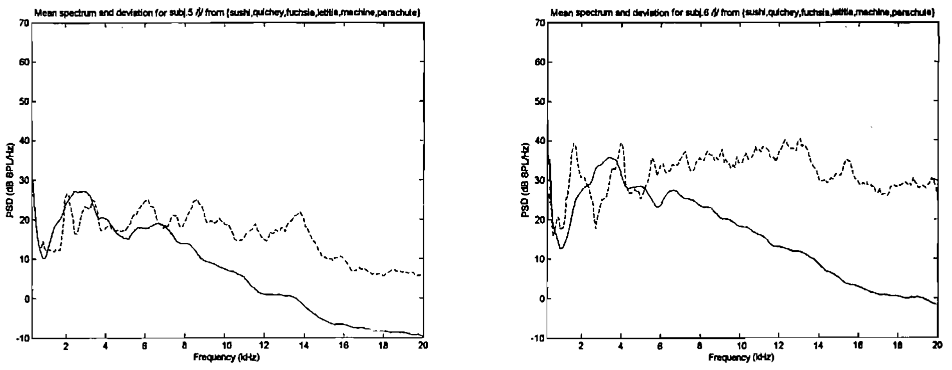


FIGURE C.8: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject M-05 (left) and M-06 (right).

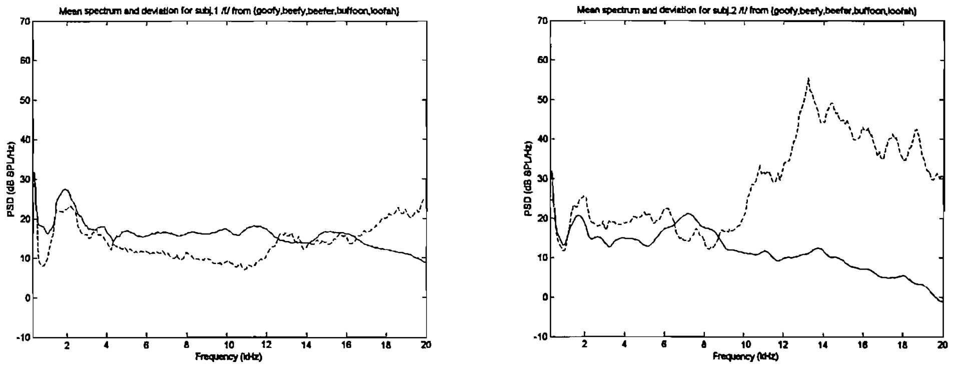


FIGURE C.9: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject M-01 (left) and M-02 (right).

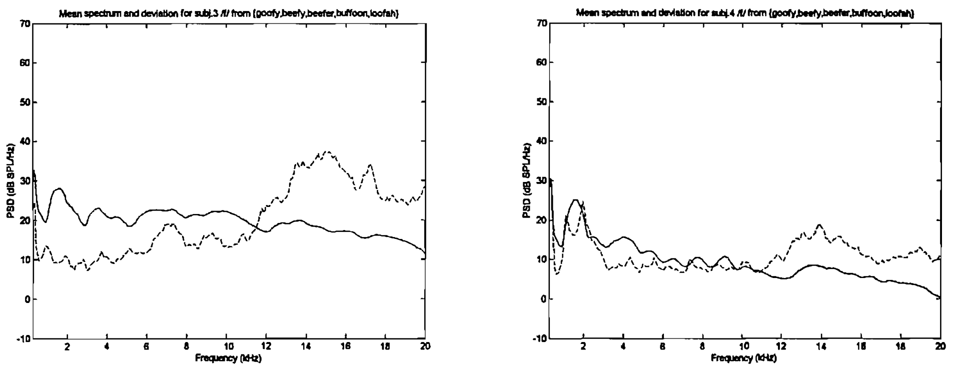


FIGURE C.10: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject M-03 (left) and M-04 (right).

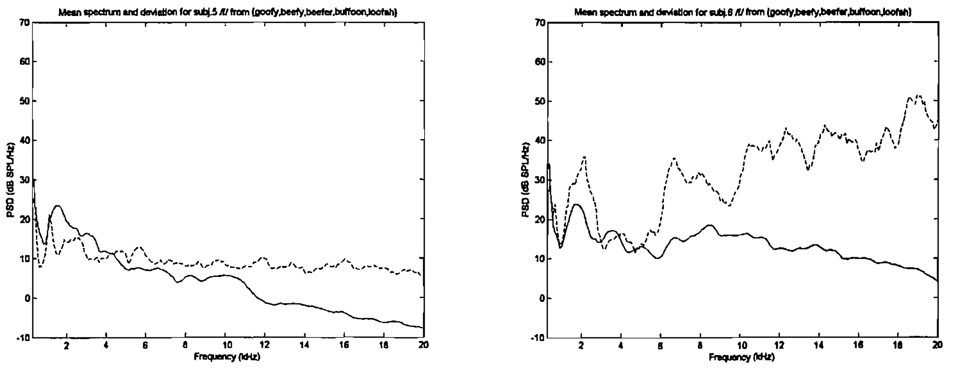


FIGURE C.11: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject M-05 (left) and M-06 (right).

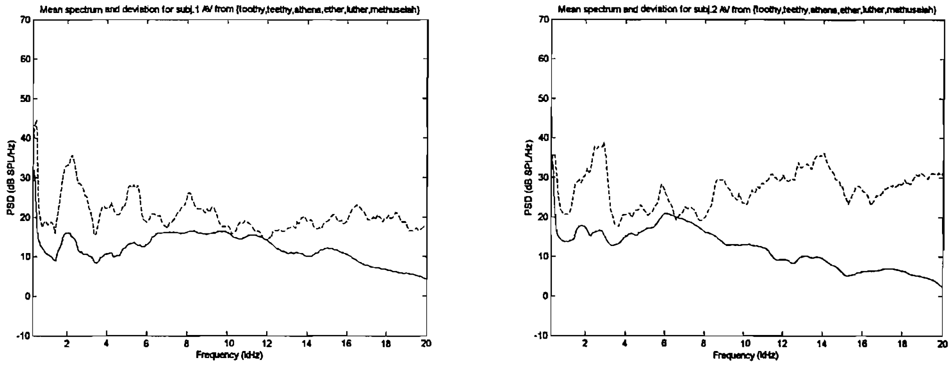


FIGURE C.12: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject M-01 (left) and M-02 (right).

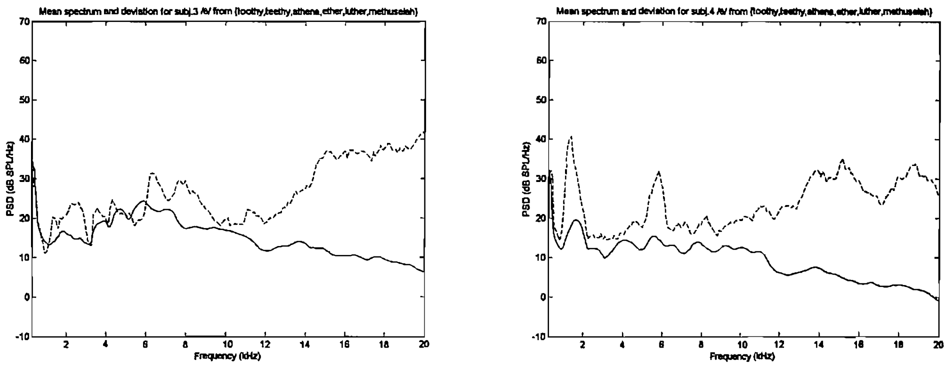


FIGURE C.13: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject M-03 (left) and M-04 (right).

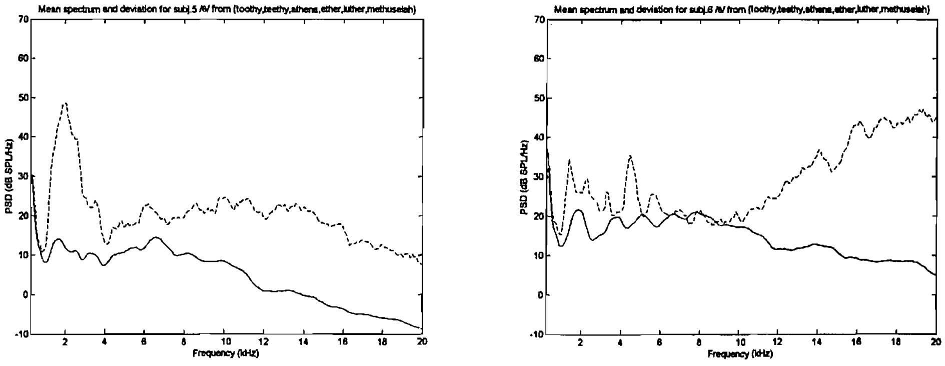


FIGURE C.14: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject M-05 (left) and M-06 (right).

## C.2 Females

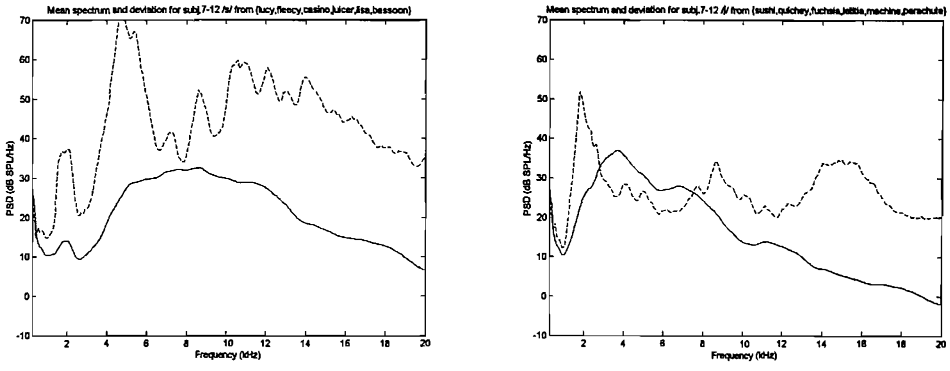


FIGURE C.15: Mean spectrum (solid) and production variance (dashed) from central /s/ (left) in /usi,isi,əsi,isə,əsu,usə/ contexts, and /ʃ/ (right) in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, all female subjects.

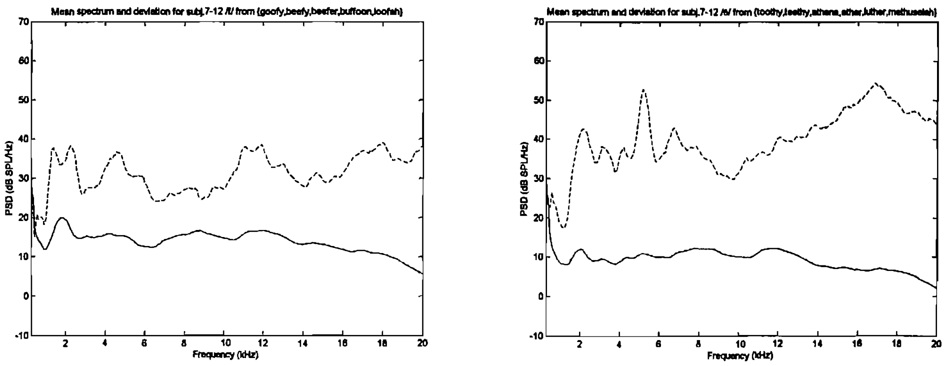


FIGURE C.16: Mean spectrum (solid) and production variance (dashed) from central /f/ (left) in /ufi,ifi,ifə,əfu,ufə/ contexts, and /θ/ (right) in /uθi,iθi,əθi,iθə,uθə,əθu/ contexts, all female subjects.

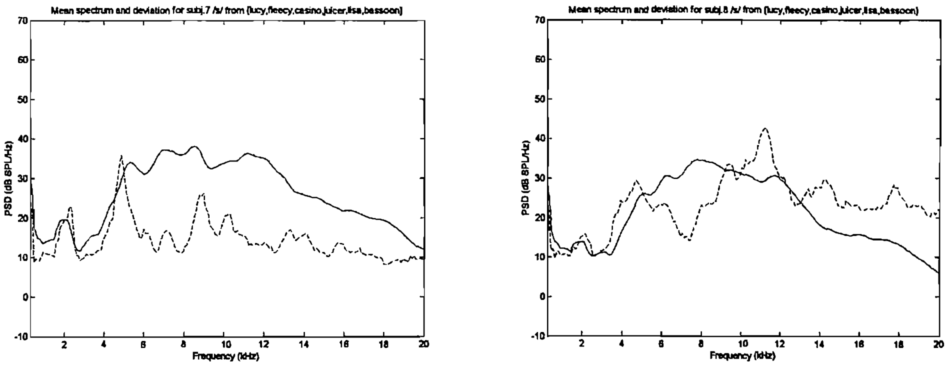


FIGURE C.17: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject F-07 (left) and F-08 (right).

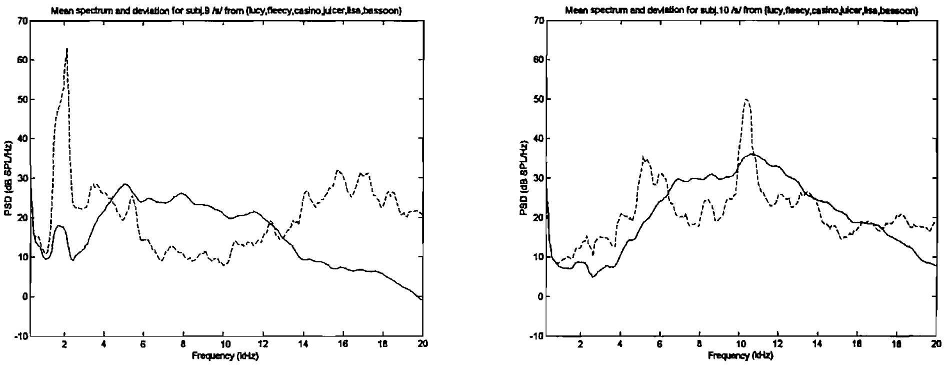


FIGURE C.18: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject F-09 (left) and F-10 (right).

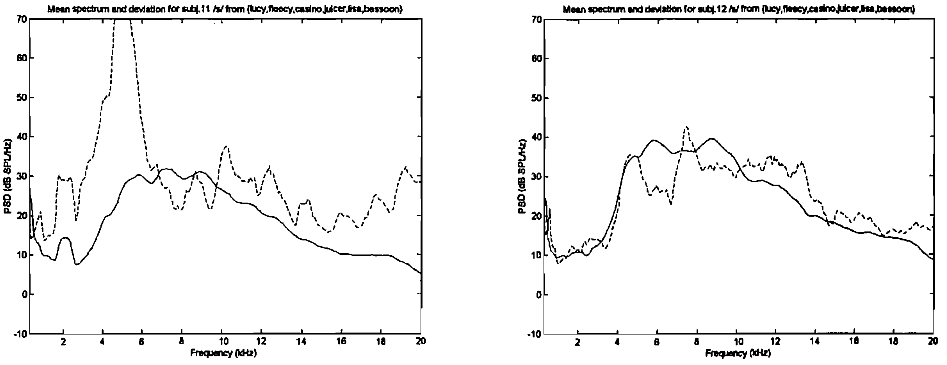


FIGURE C.19: Mean spectrum (solid) and production variance (dashed) from central /s/ in /usi,isi,əsi,isə,əsu,usə/ contexts, subject F-11 (left) and F-12 (right).

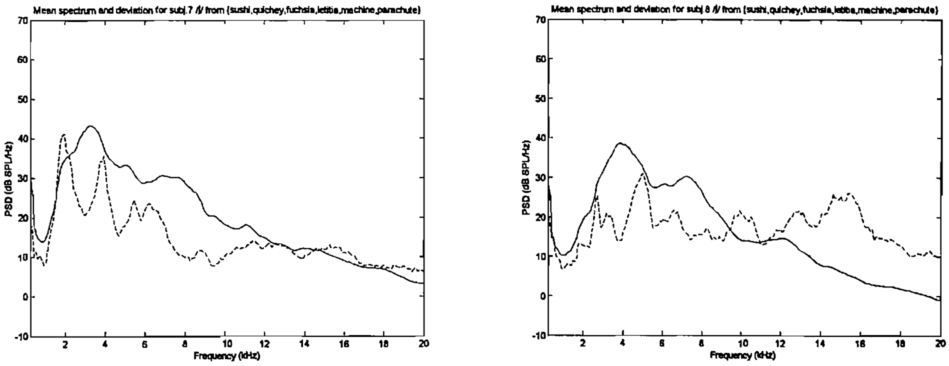


FIGURE C.20: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject F-07 (left) and F-08 (right).

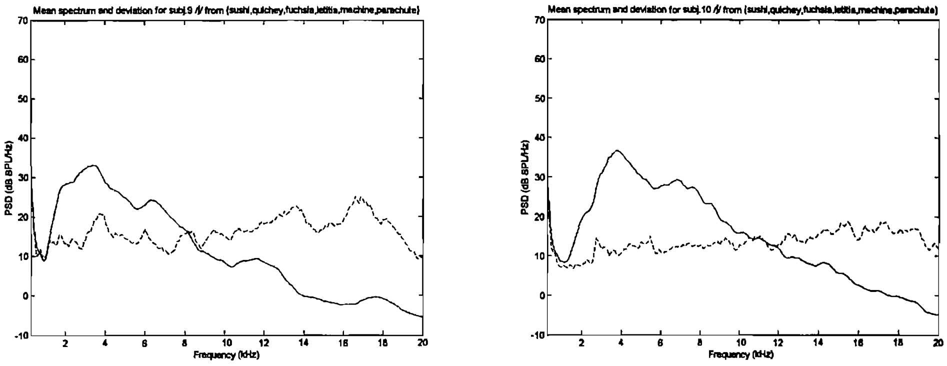


FIGURE C.21: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject F-09 (left) and F-10 (right).

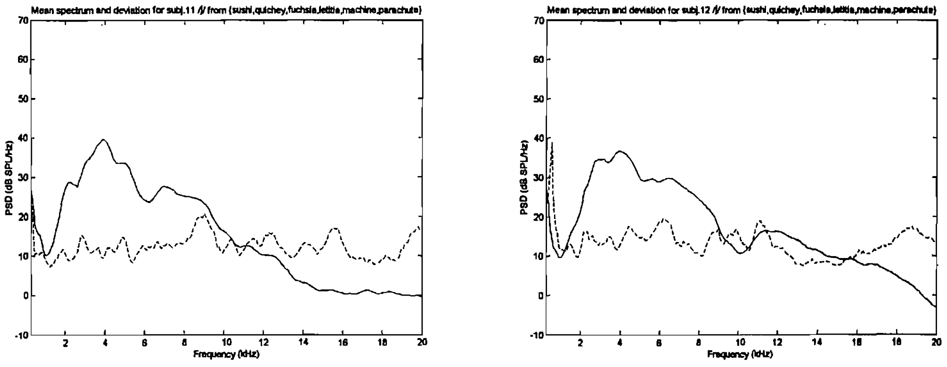


FIGURE C.22: Mean spectrum (solid) and production variance (dashed) from central /ʃ/ in /uʃi,iʃi,əʃi,iʃə,əʃu,uʃə/ contexts, subject F-11 (left) and F-12 (right).

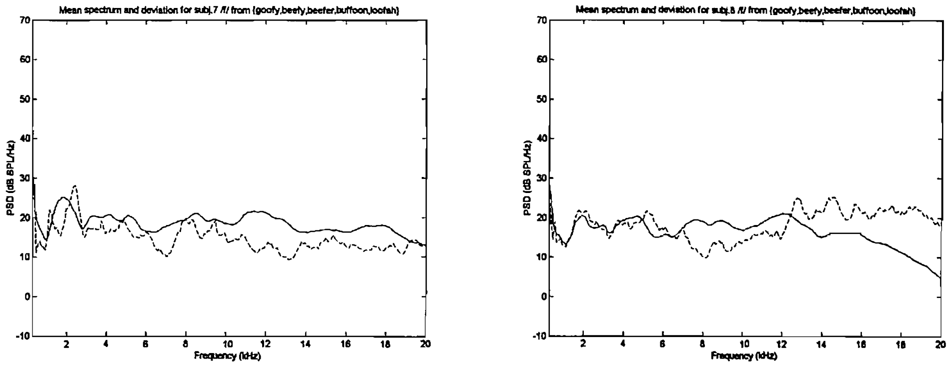


FIGURE C.23: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject F-07 (left) and F-08 (right).

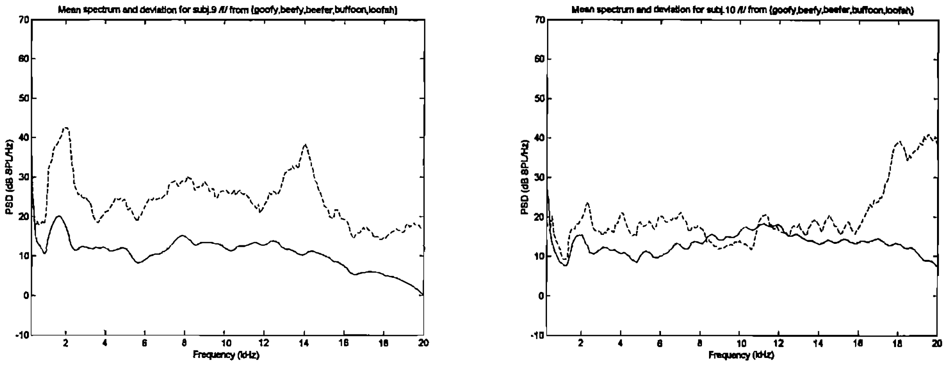


FIGURE C.24: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject F-09 (left) and F-10 (right).

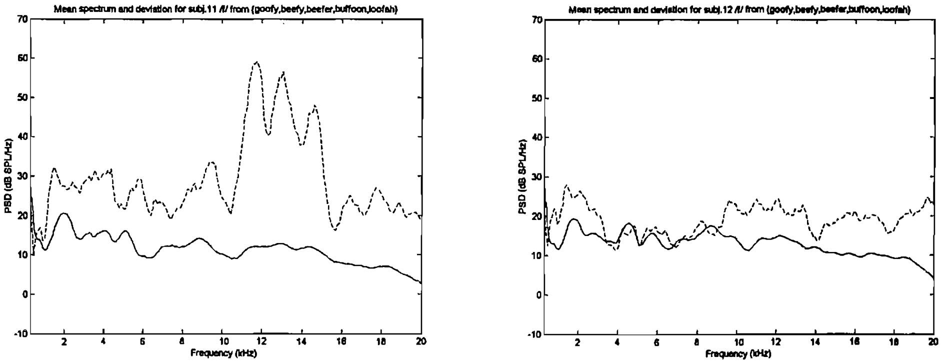


FIGURE C.25: Mean spectrum (solid) and production variance (dashed) from central /f/ in /ufi,ifi,ifə,əfu,ufə/ contexts, subject F-11 (left) and F-12 (right).

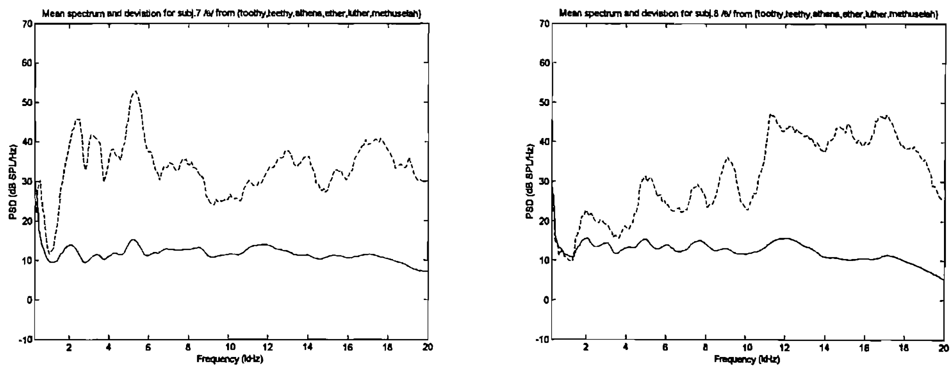


FIGURE C.26: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject F-07 (left) and F-08 (right).

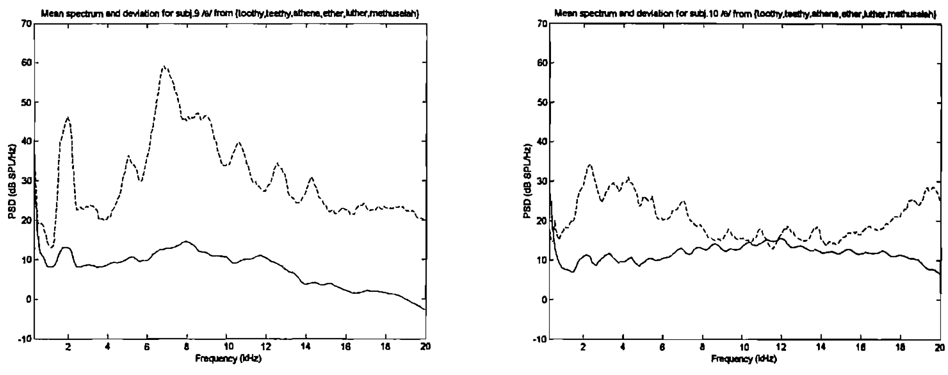


FIGURE C.27: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject F-09 (left) and F-10 (right).

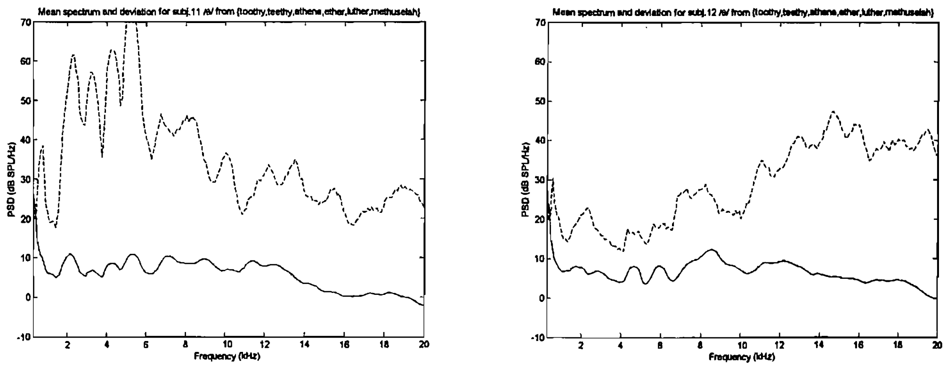


FIGURE C.28: Mean spectrum (solid) and production variance (dashed) from central /θ/ in /uθi,iθi,iθə,əθu,uθə,əθi/ contexts, subject F-11 (left) and F-12 (right).

# Appendix D

# Coefficients of correlation of amplitude to spectrum in fricative centres

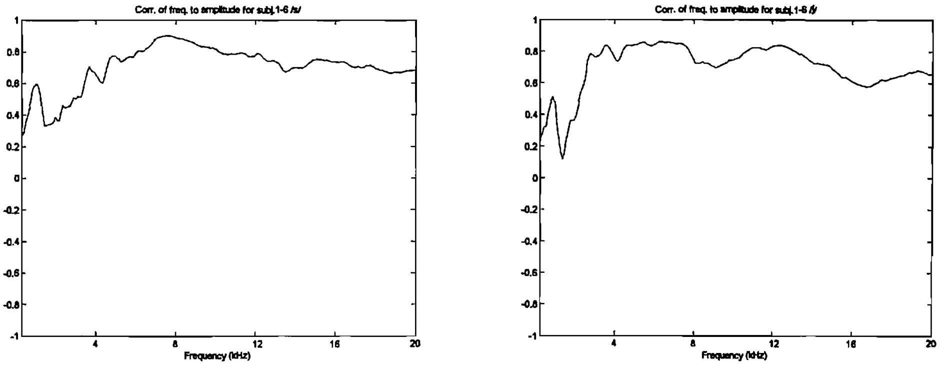


FIGURE D.1: Coefficients of correlation of amplitude to spectrum in central /s/ and /ʃ/ spectra, across all subjects.

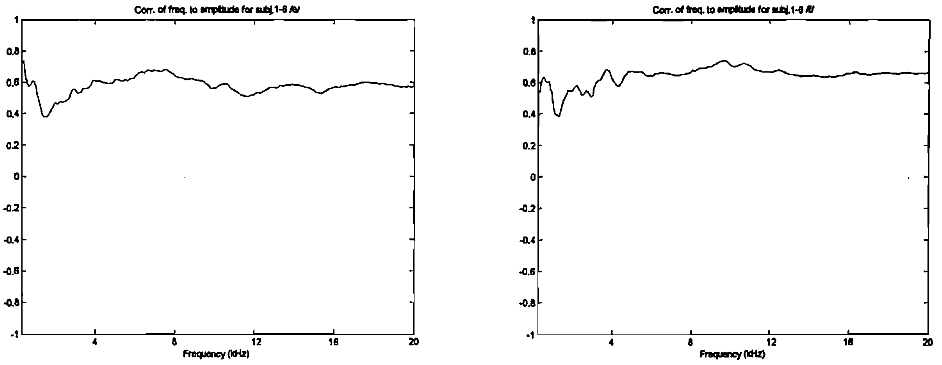


FIGURE D.2: Coefficients of correlation of amplitude to spectrum in central /θ/ and /f/ spectra, across all subjects.

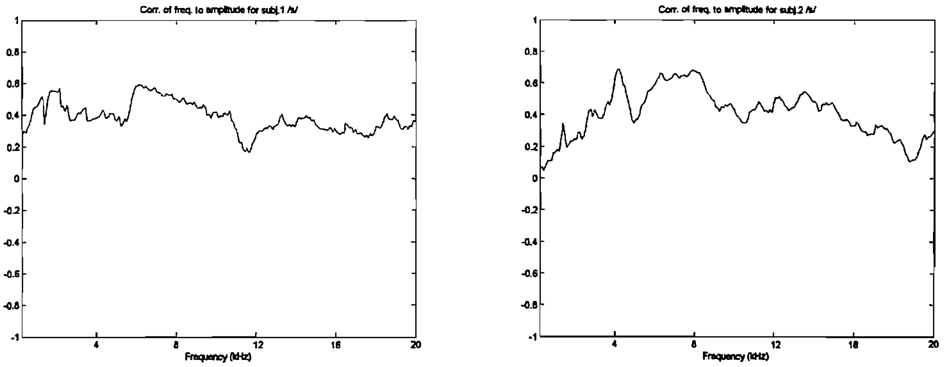


FIGURE D.3: Coefficients of correlation of amplitude to spectrum of 6 central /s/ spectra, subjects M-01 and M-02.

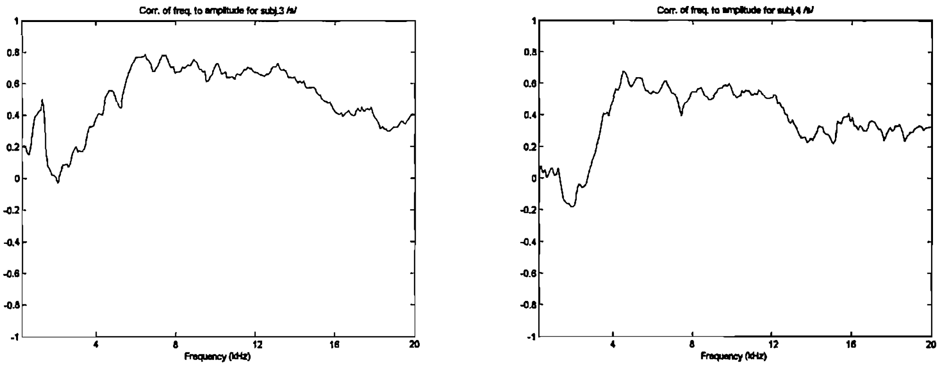


FIGURE D.4: Coefficients of correlation of amplitude to spectrum of 6 central /s/ spectra, subjects M-03 and M-04.

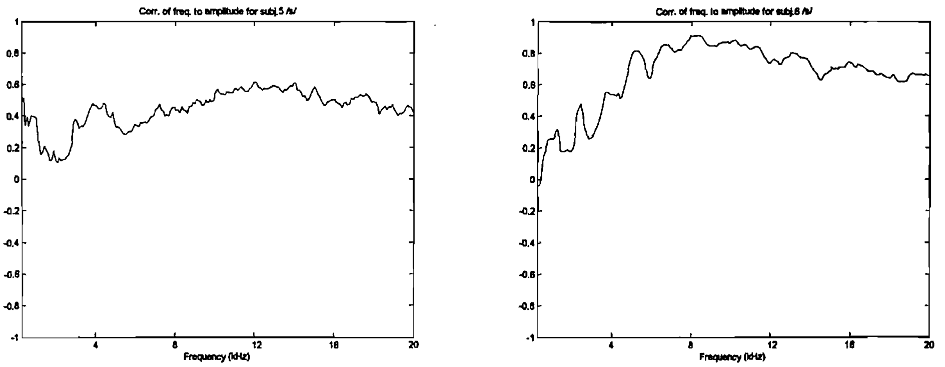


FIGURE D.5: Coefficients of correlation of amplitude to spectrum of 6 central /s/ spectra, subjects M-05 and M-06.

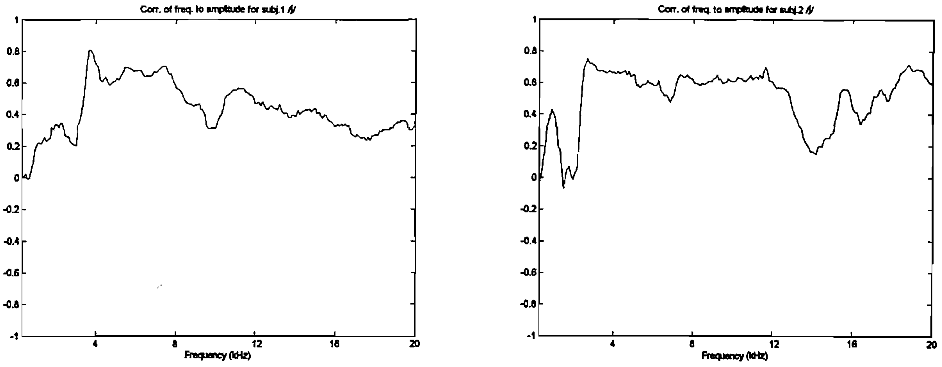


FIGURE D.6: Coefficients of correlation of amplitude to spectrum of central /ʃ/ spectra, subjects M-01 and M-02.

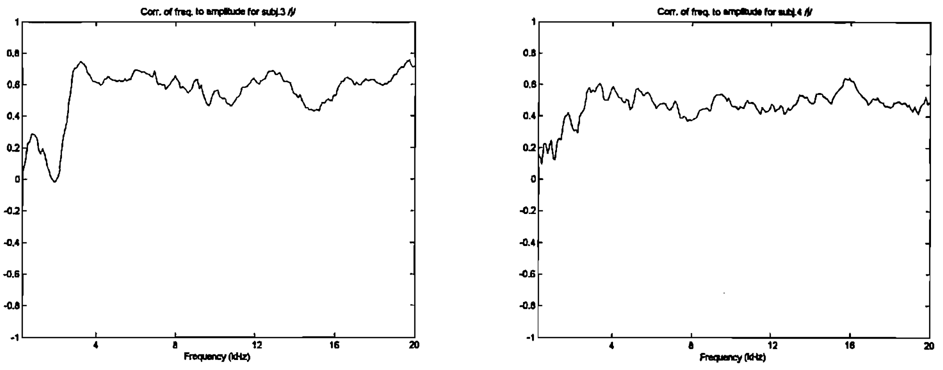


FIGURE D.7: Coefficients of correlation of amplitude to spectrum of central /ʃ/ spectra, subjects M-03 and M-04.

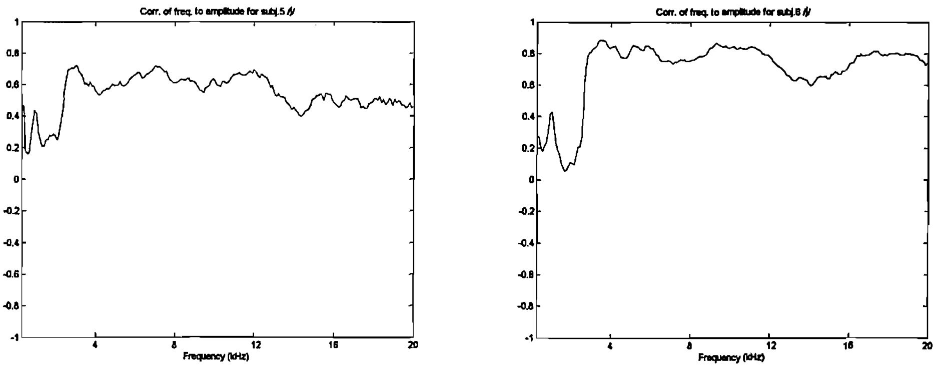


FIGURE D.8: Coefficients of correlation of amplitude to spectrum of central /ʃ/ spectra, subjects M-05 and M-06.

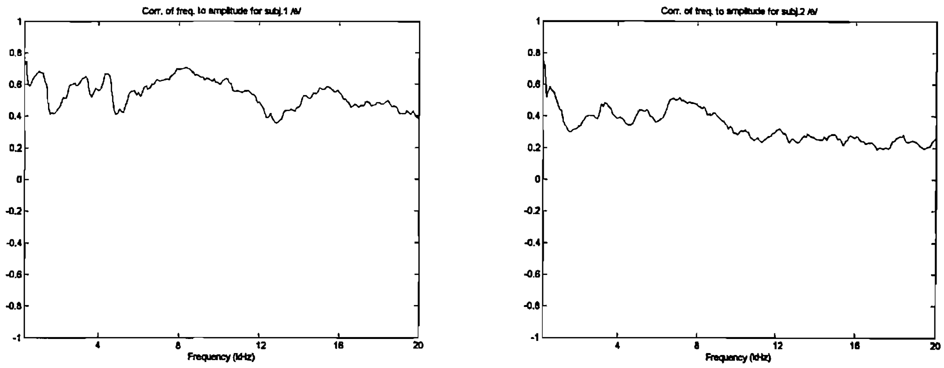


FIGURE D.9: Coefficients of correlation of amplitude to spectrum of central /θ/ spectra, subjects M-01 and M-02.

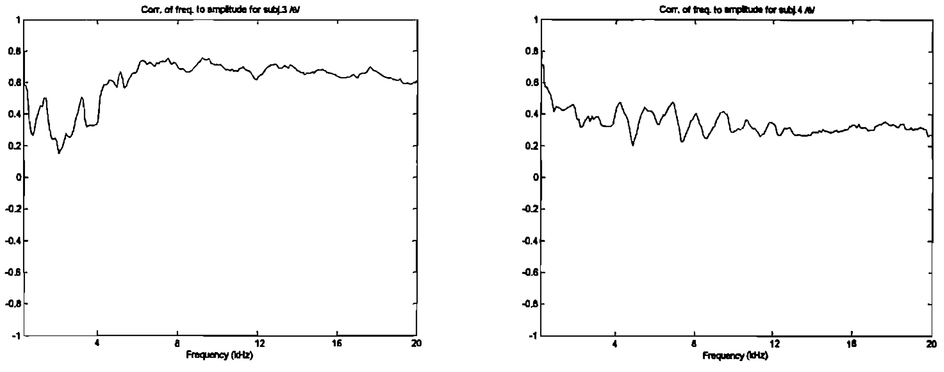


FIGURE D.10: Coefficients of correlation of amplitude to spectrum of central /θ/ spectra, subjects M-03 and M-04.

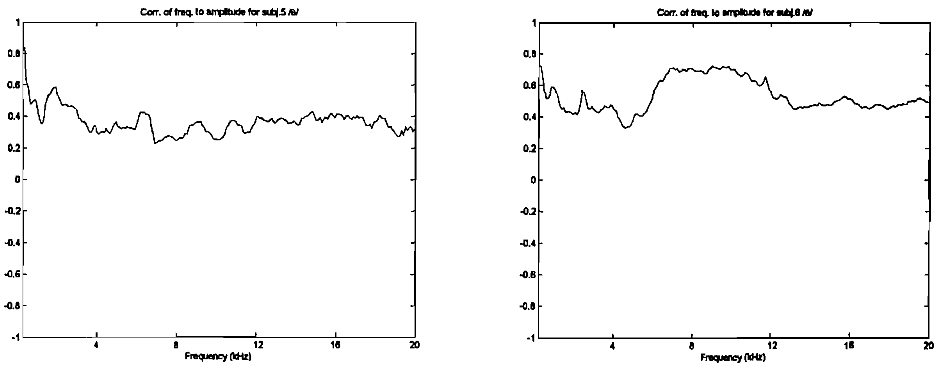


FIGURE D.11: Coefficients of correlation of amplitude to spectrum of central /θ/ spectra, subjects M-05 and M-06.

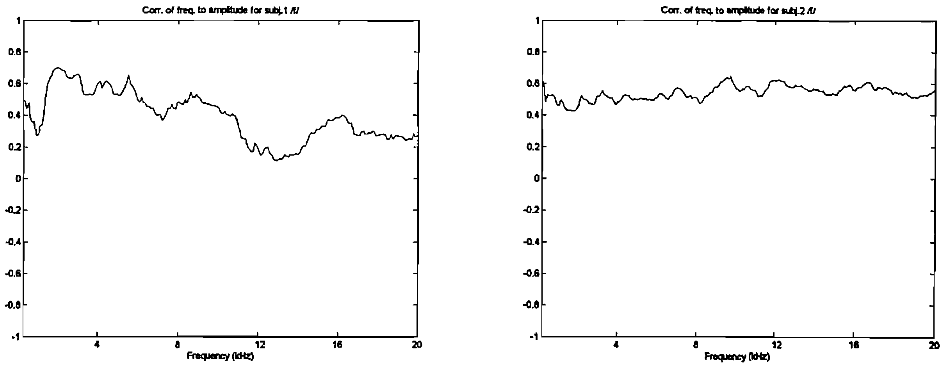


FIGURE D.12: Coefficients of correlation of amplitude to spectrum of central /f/ spectra, subjects M-01 and M-02.

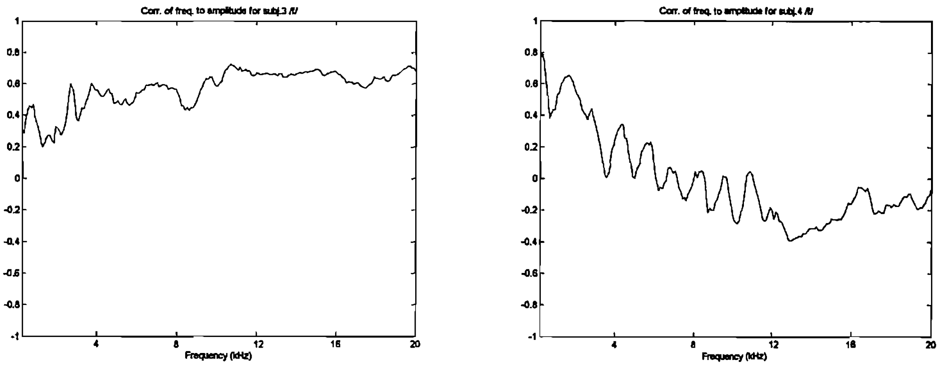


FIGURE D.13: Coefficients of correlation of amplitude to spectrum of central /f/ spectra, subjects M-03 and M-04.

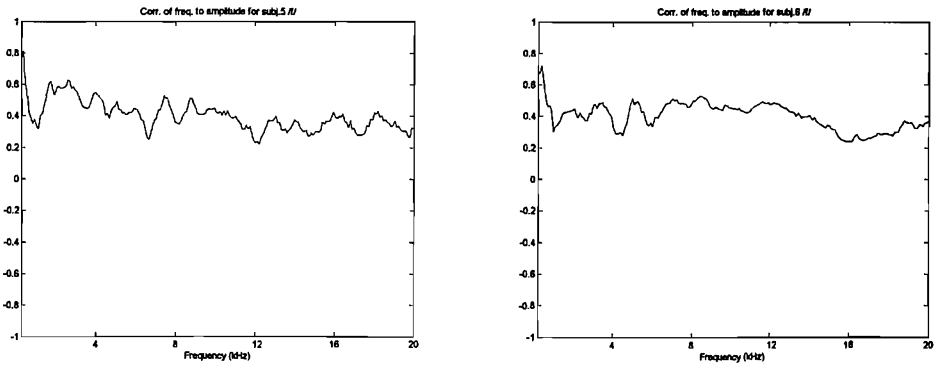


FIGURE D.14: Coefficients of correlation of amplitude to spectrum of central /f/ spectra, subjects M-05 and M-06.

# Appendix E

# Coefficients of spectral correlation in fricative centres, male tokens

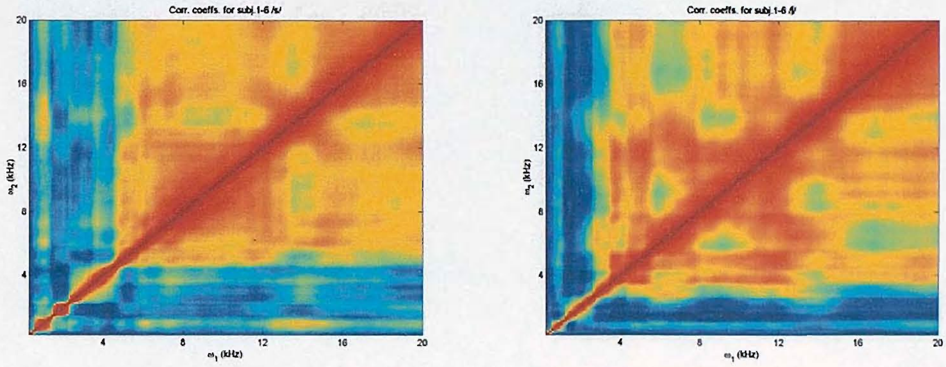


FIGURE E.1: Coefficients of correlation of central /s/ (left) and /ʃ/ (right) spectra, across all male subjects.

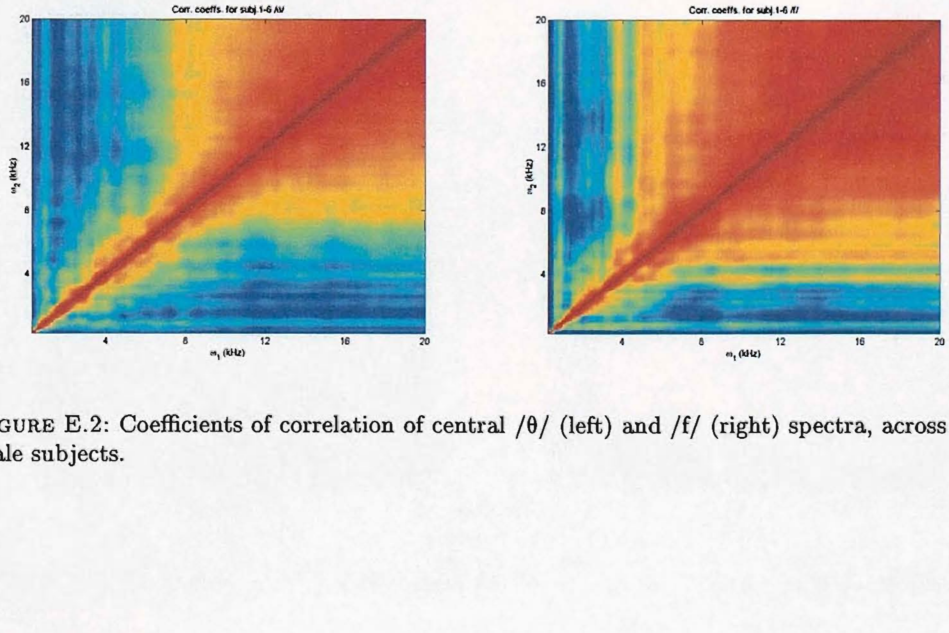


FIGURE E.2: Coefficients of correlation of central /θ/ (left) and /f/ (right) spectra, across all male subjects.

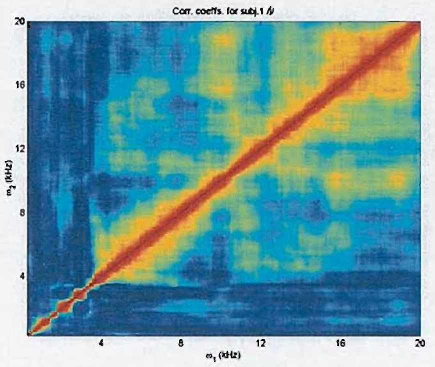


FIGURE E.3: Coefficients of spectral correlation of central /ʃ/ spectra, subjects M-01 and M-02.

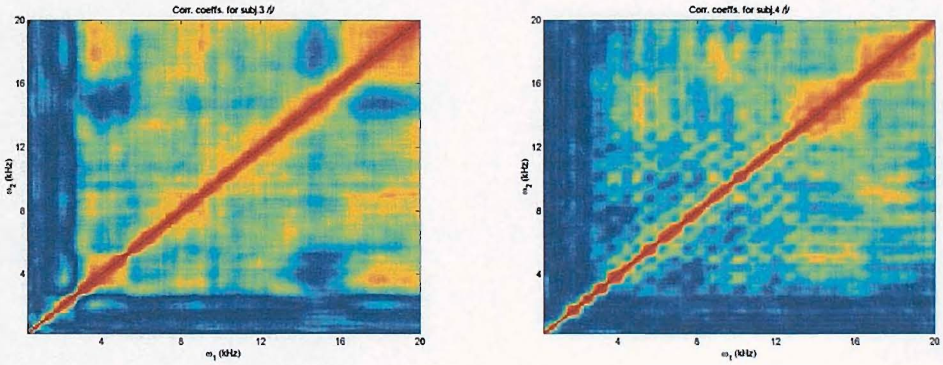


FIGURE E.4: Coefficients of spectral correlation of central /ʃ/ spectra, subjects M-03 and M-04.

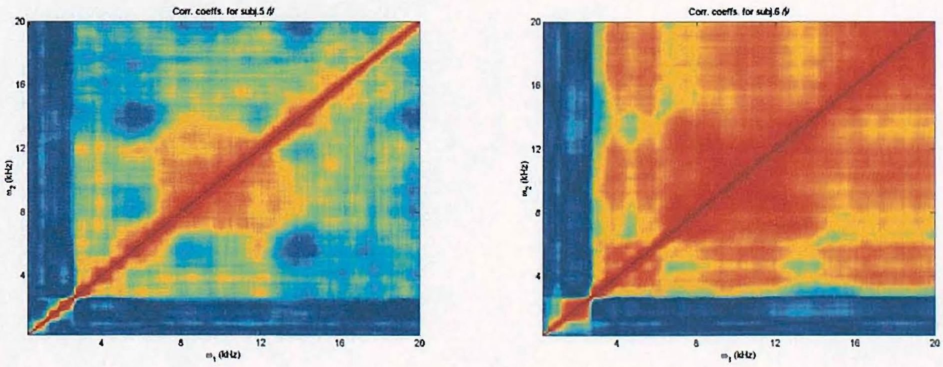


FIGURE E.5: Coefficients of spectral correlation of central /ʃ/ spectra, subjects M-05 and M-06.

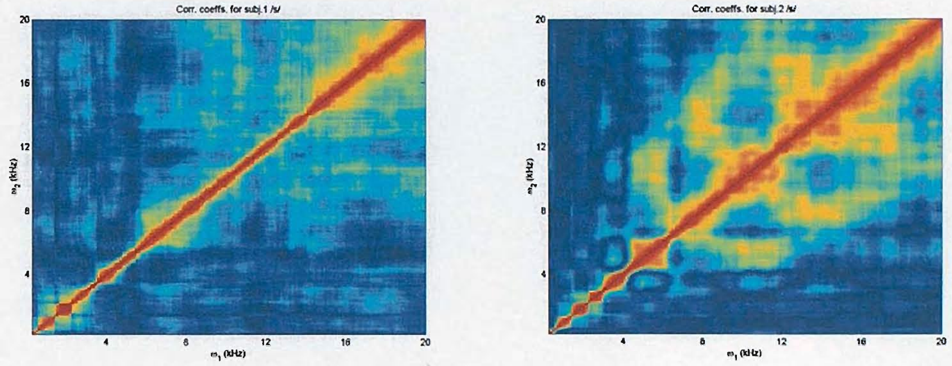


FIGURE E.6: Coefficients of spectral correlation of central /s/ spectra, subjects M-01 and M-02.

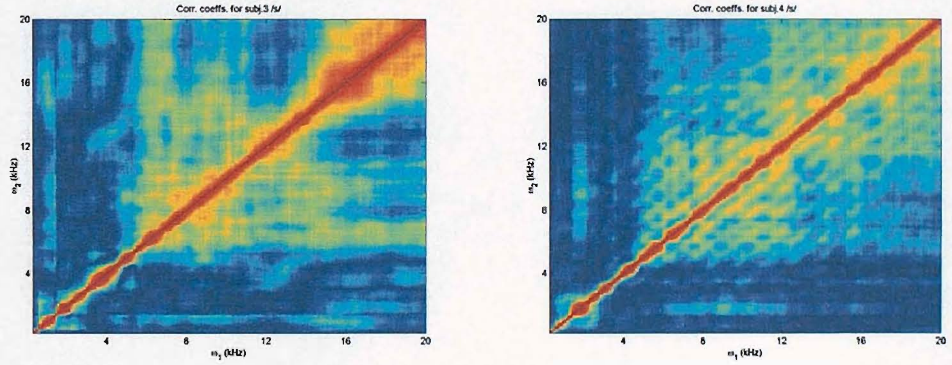


FIGURE E.7: Coefficients of spectral correlation of central /s/ spectra, subjects M-03 and M-04.

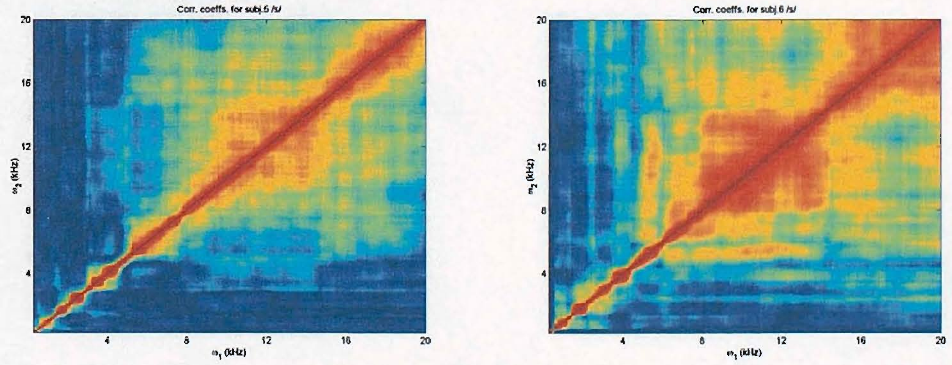


FIGURE E.8: Coefficients of spectral correlation of central /s/ spectra, subjects M-05 and M-06.

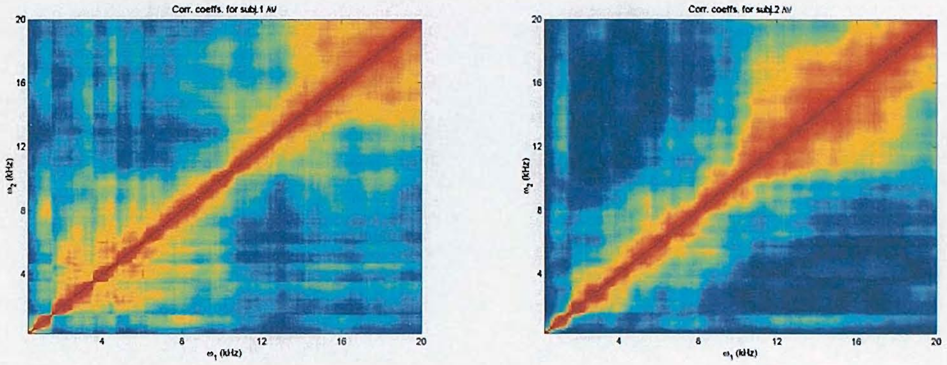


FIGURE E.9: Coefficients of spectral correlation of central /θ/ spectra, subjects M-01 and M-02.

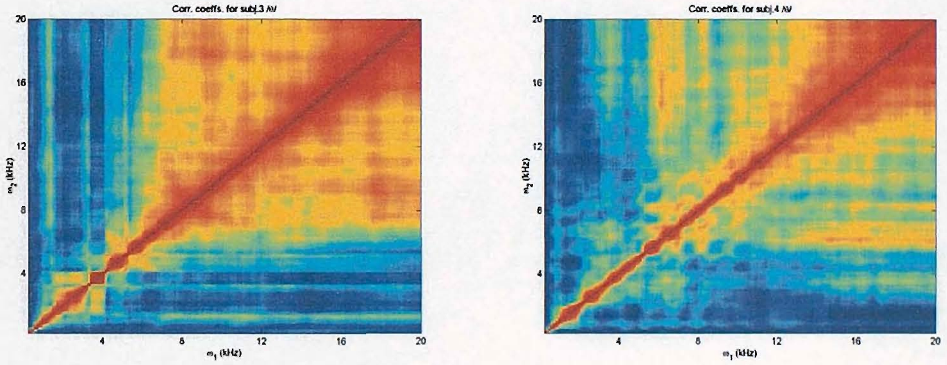


FIGURE E.10: Coefficients of spectral correlation of central /θ/ spectra, subjects M-03 and M-04.

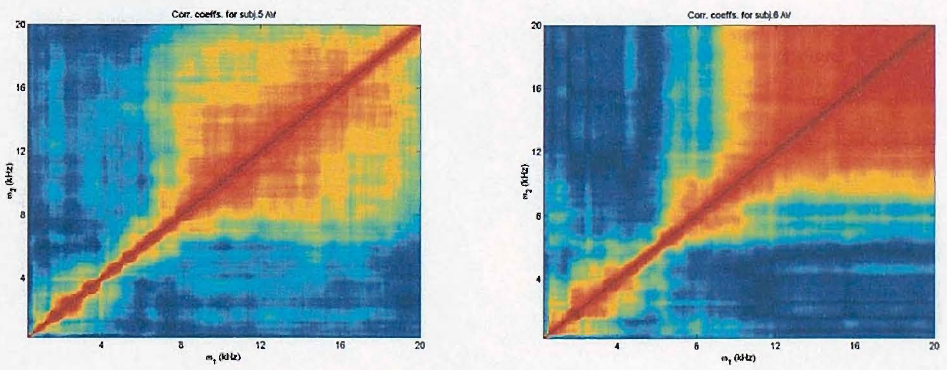


FIGURE E.11: Coefficients of spectral correlation of central /θ/ spectra, subjects M-05 and M-06.

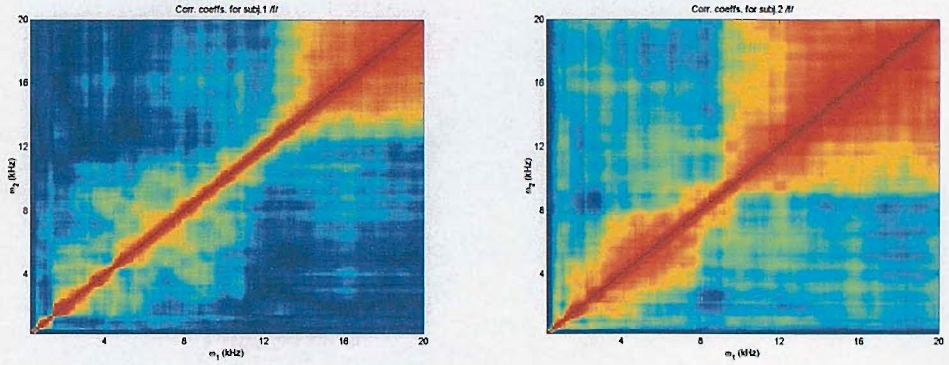


FIGURE E.12: Coefficients of spectral correlation of central /f/ spectra, subjects M-01 and M-02.

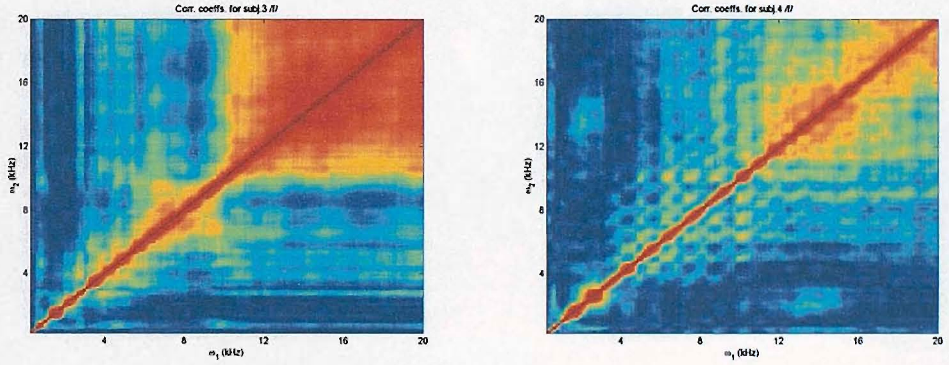


FIGURE E.13: Coefficients of spectral correlation of central /f/ spectra, subjects M-03 and M-04.

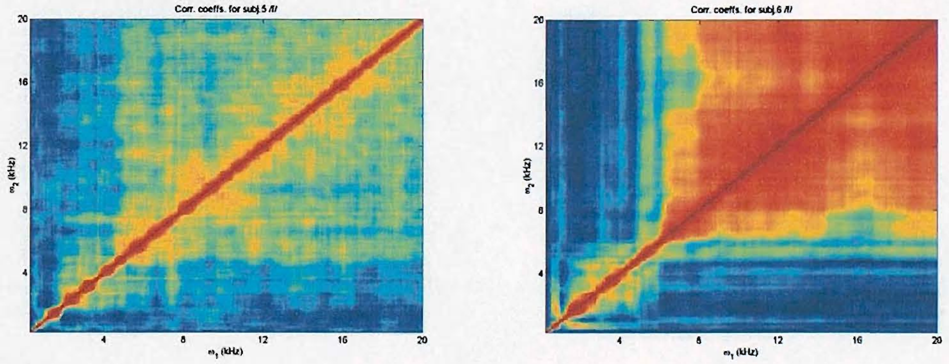


FIGURE E.14: Coefficients of spectral correlation of central /f/ spectra, subjects M-05 and M-06.

# Appendix F

# Within vowel-context measurements, by speaker

In all graphs, /s/ tokens are represented by a blue cross, /ʃ/ by a red circle, /f/ by a green cross, and /θ/ by a black dot.

## F.1 1st and 2nd spectral moments by subject and example vowel context

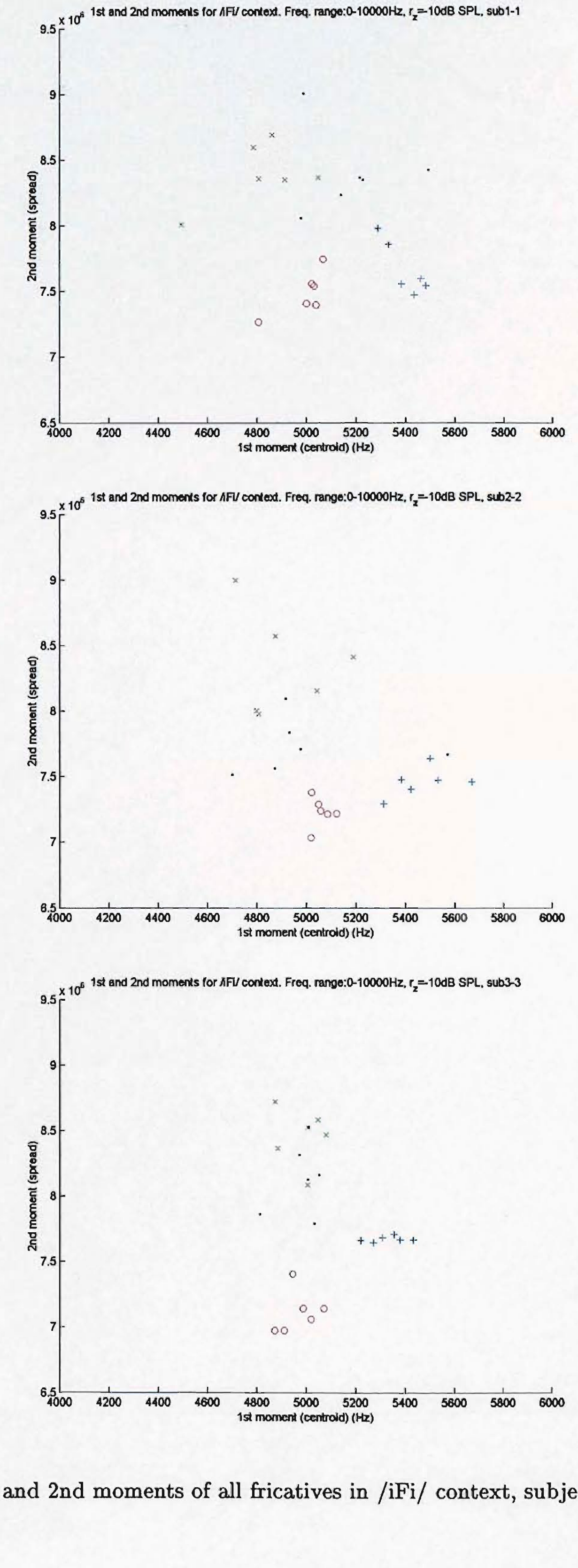


FIGURE F.1: 1st and 2nd moments of all fricatives in /iFi/ context, subjects M-01 to M-03.

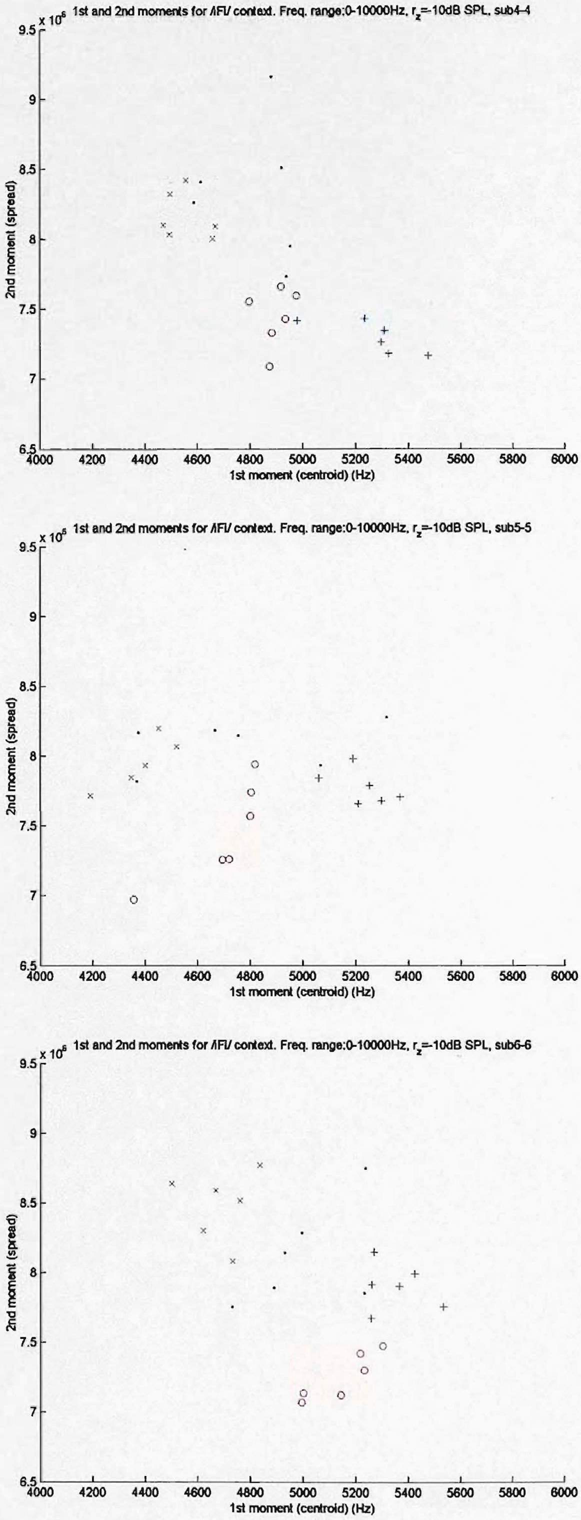


FIGURE F.2: 1st and 2nd moments of all fricatives in /iFi/ context, subjects M-04 to M-06.

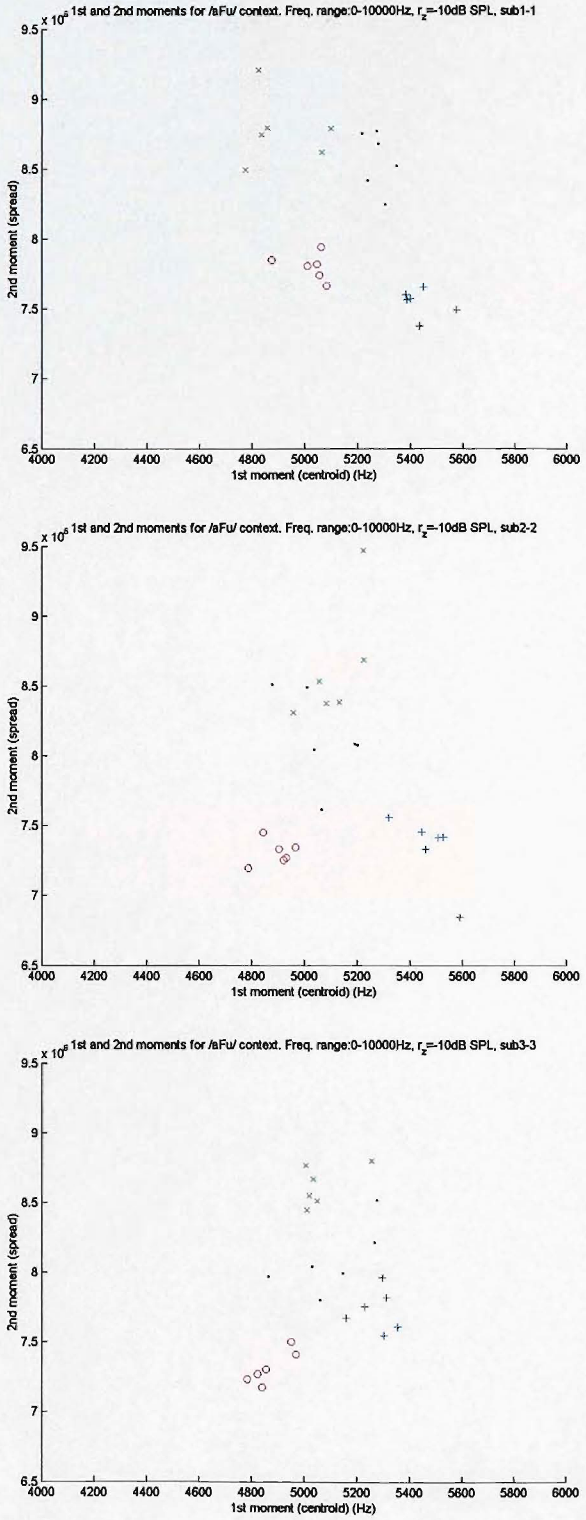


FIGURE F.3: 1st and 2nd moments of all fricatives in /əFu/ context, subjects M-01 to M-03.

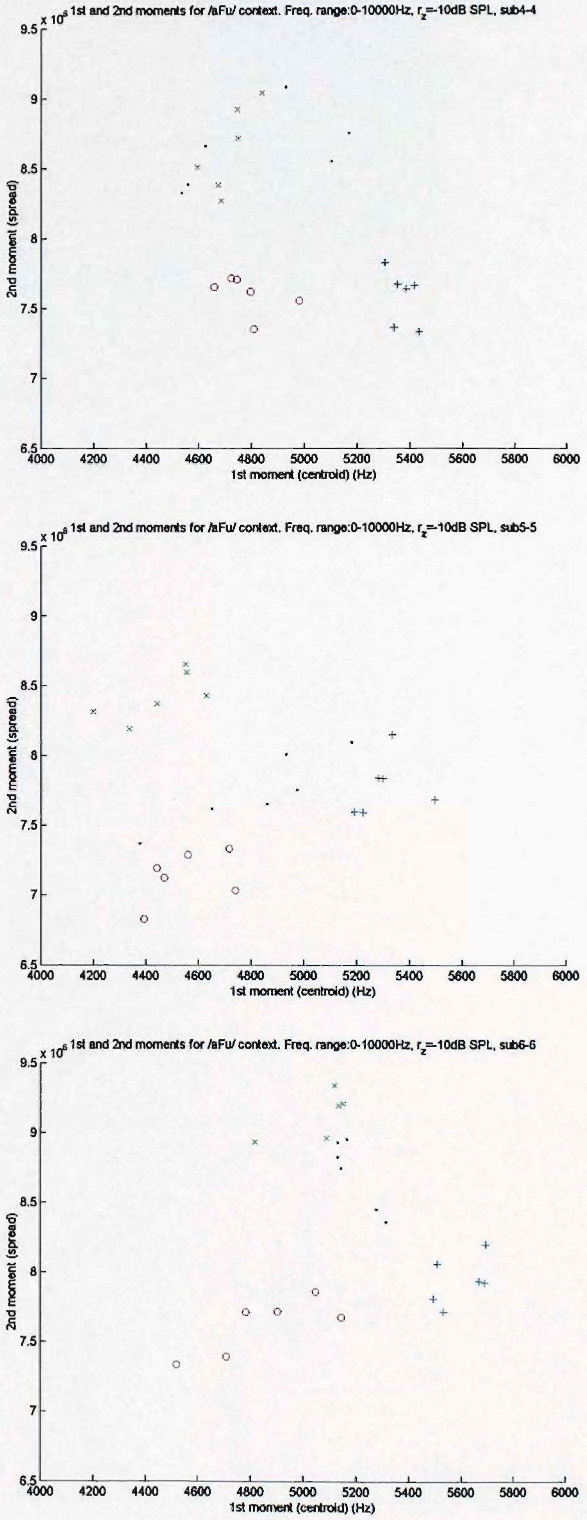


FIGURE F.4: 1st and 2nd moments of all fricatives in /əFu/ context, subjects M-04 to M-06.

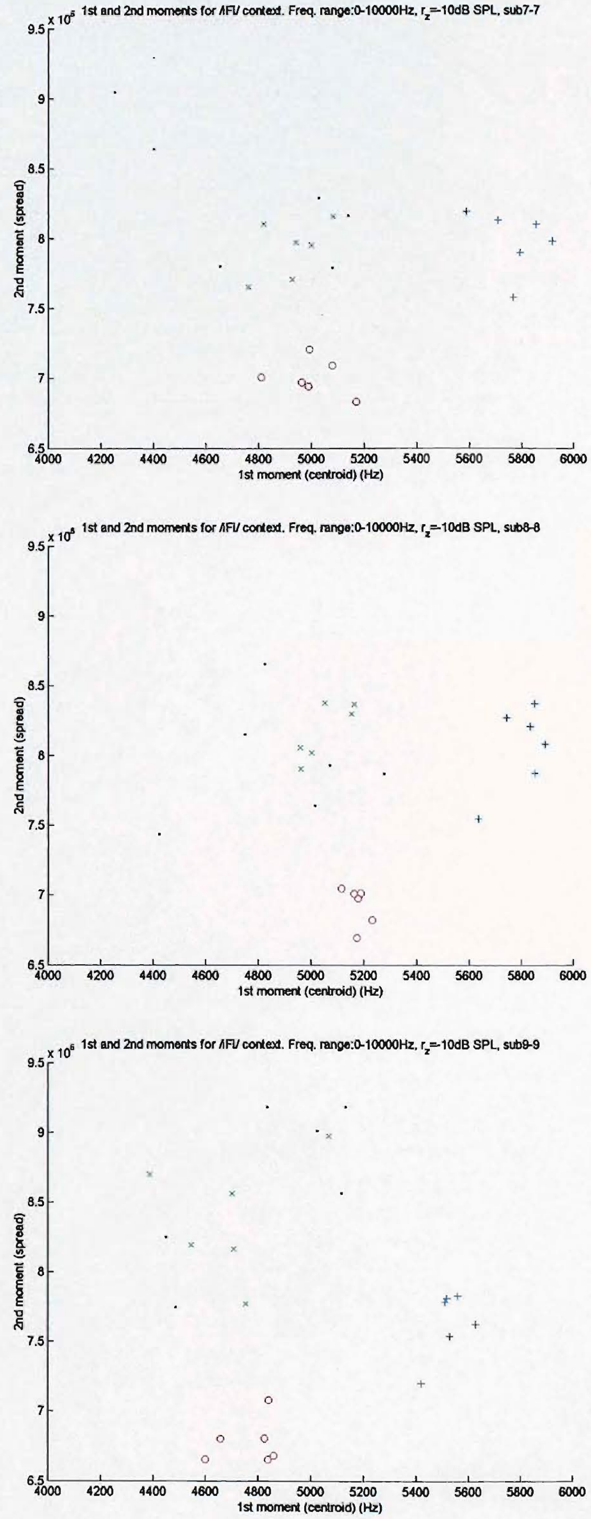


FIGURE F.5: 1st and 2nd moments of all fricatives in /iFi/ context, subjects F-07 to F-09.

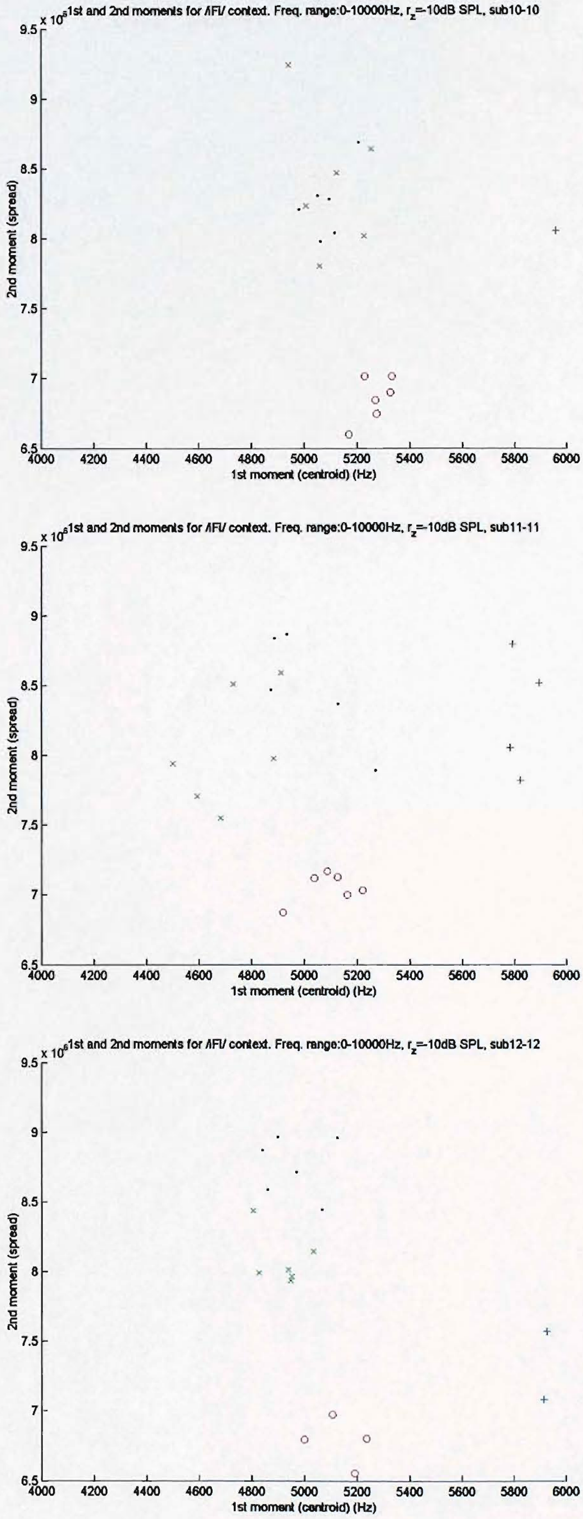


FIGURE F.6: 1st and 2nd moments of all fricatives in /iFi/ context, subjects F-10 to F-12.

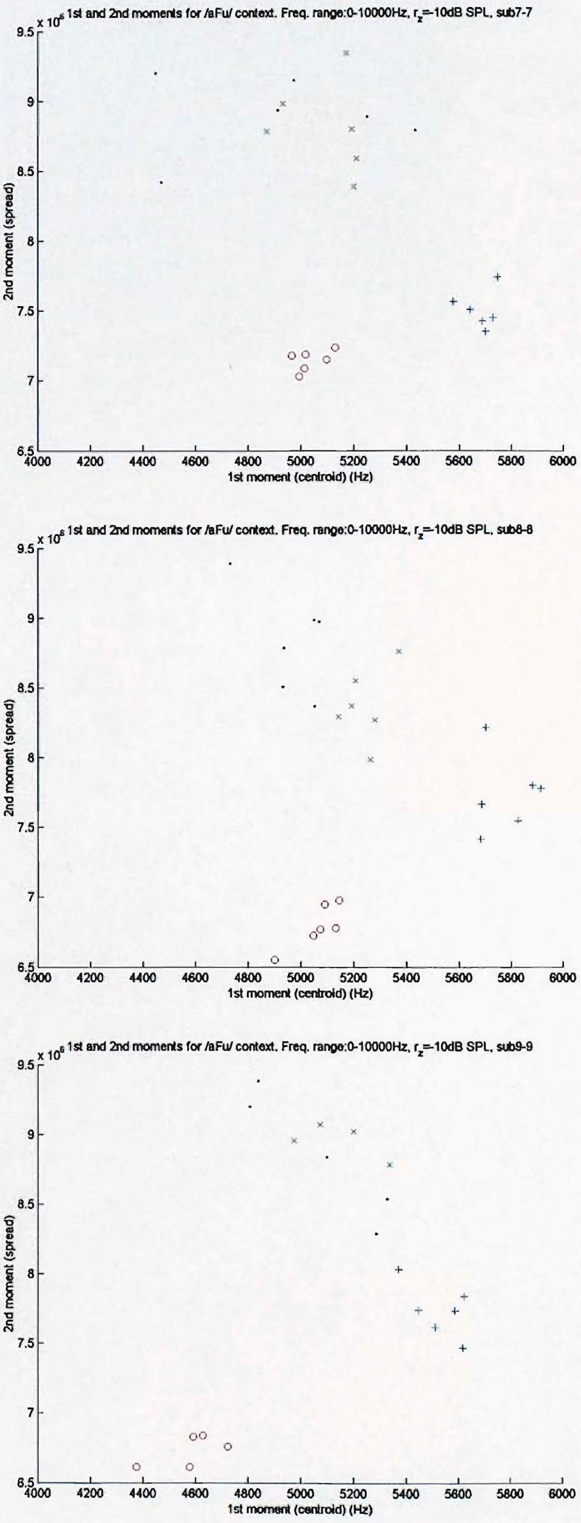


FIGURE F.7: 1st and 2nd moments of all fricatives in /əFu/ context, subjects F-07 to F-09.

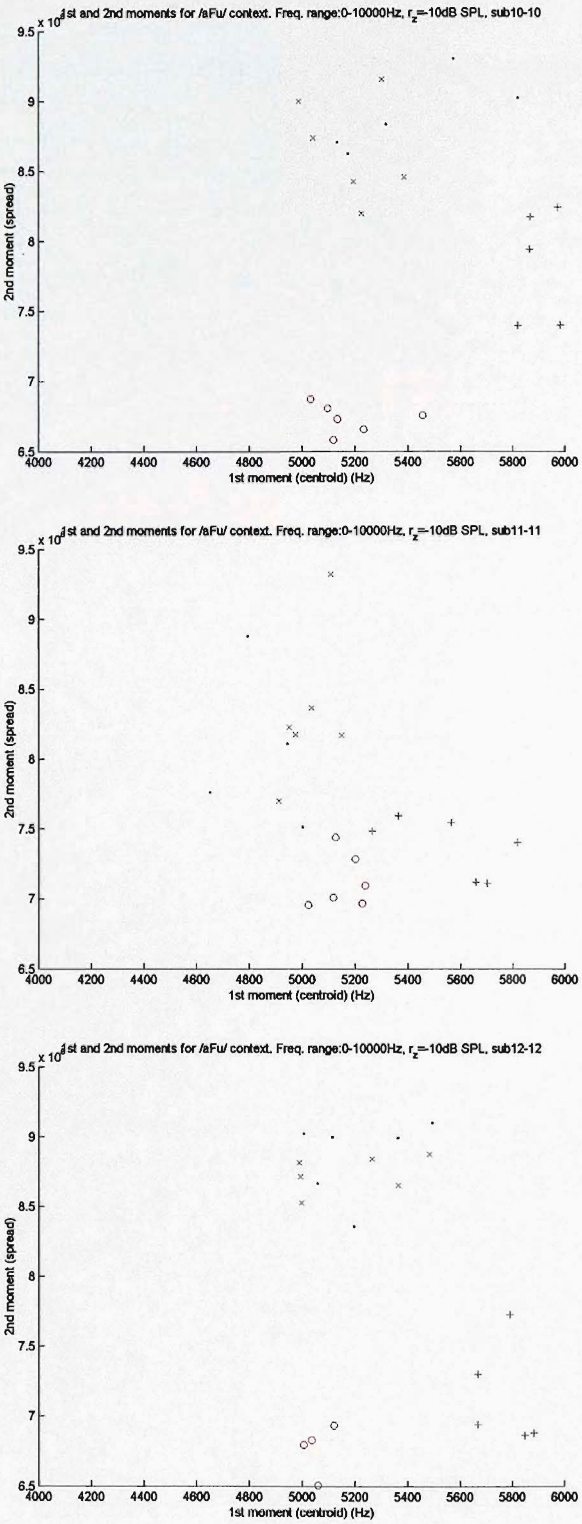


FIGURE F.8: 1st and 2nd moments of all fricatives in /əFu/ context, subjects F-10 to F-12.

## F.2 Energy measurements by subject and example vowel context

Energy at 2.5 kHz (for male tokens) and 3 kHz (for female tokens) is plotted against total spectral energy.

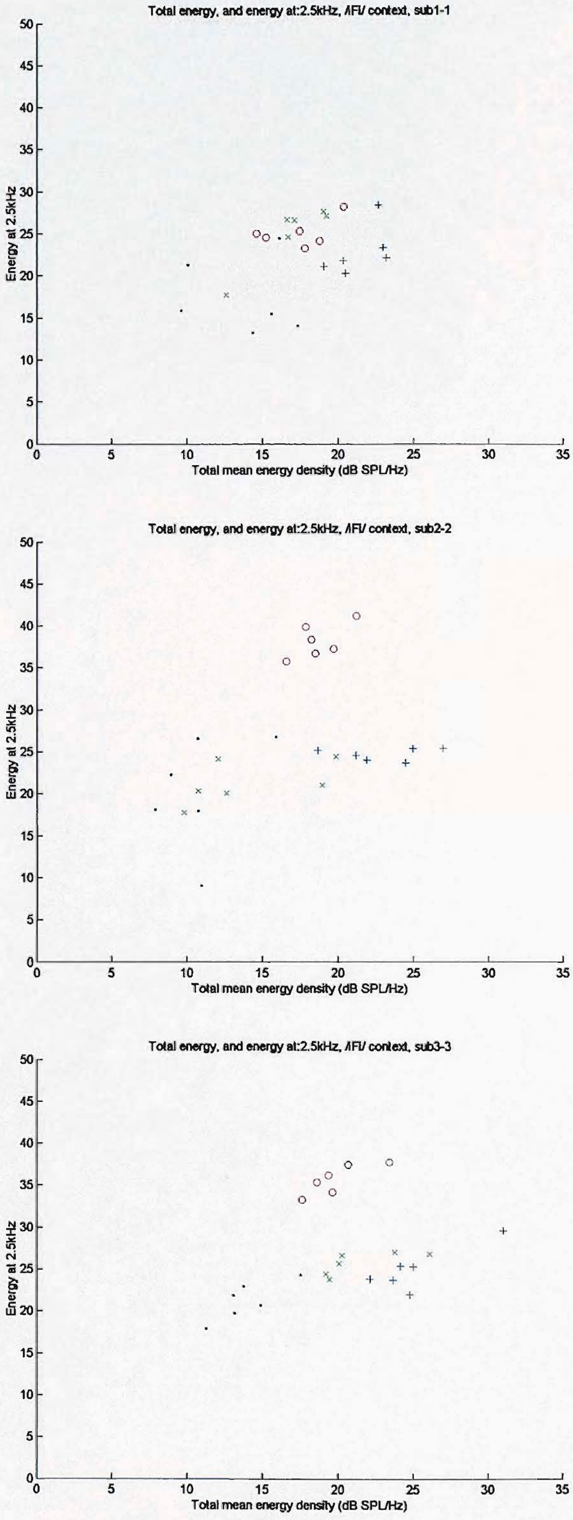
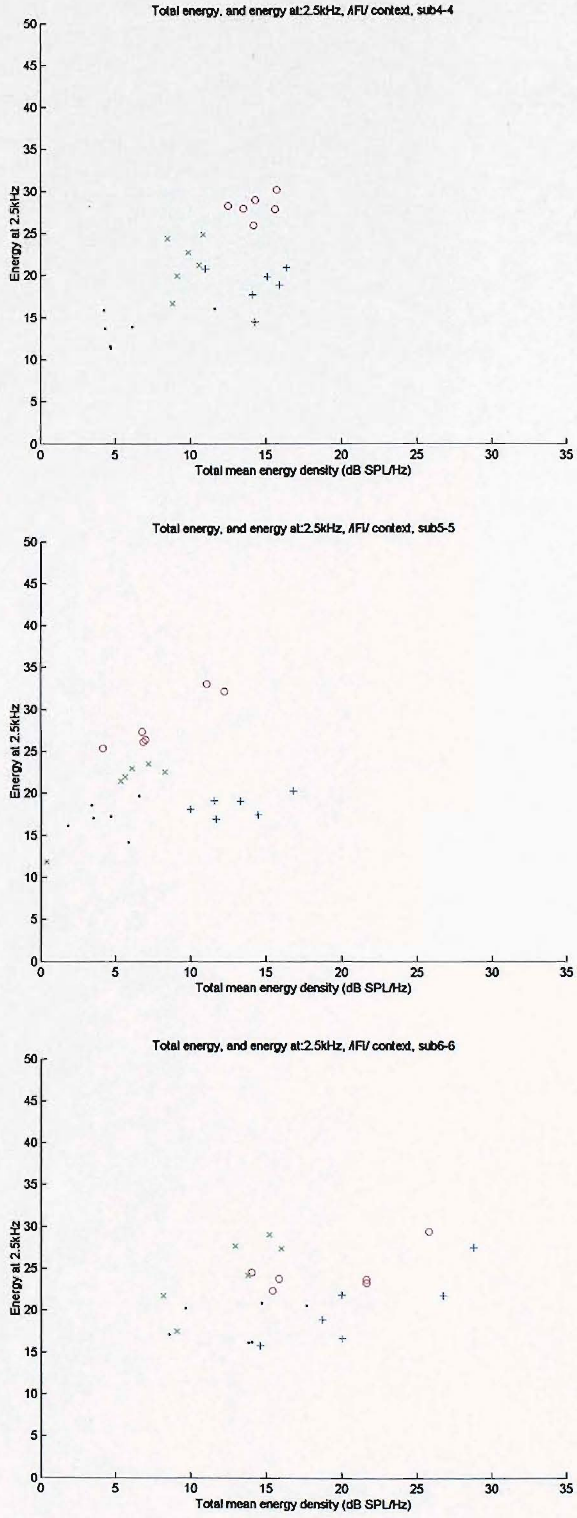


FIGURE F.9: Energy at 2.5 kHz against total spectral energy of all fricatives in /iFi/ context, subjects M-01 to M-03.



Total energy, and energy at:2.5kHz, /iFi/ context, sub4-4

Total energy, and energy at:2.5kHz, /iFi/ context, sub5-5

Total energy, and energy at:2.5kHz, /iFi/ context, sub6-6

FIGURE F.10: Energy at 2.5 kHz against total spectral energy of all fricatives in /iFi/ context, subjects M-04 to M-06.

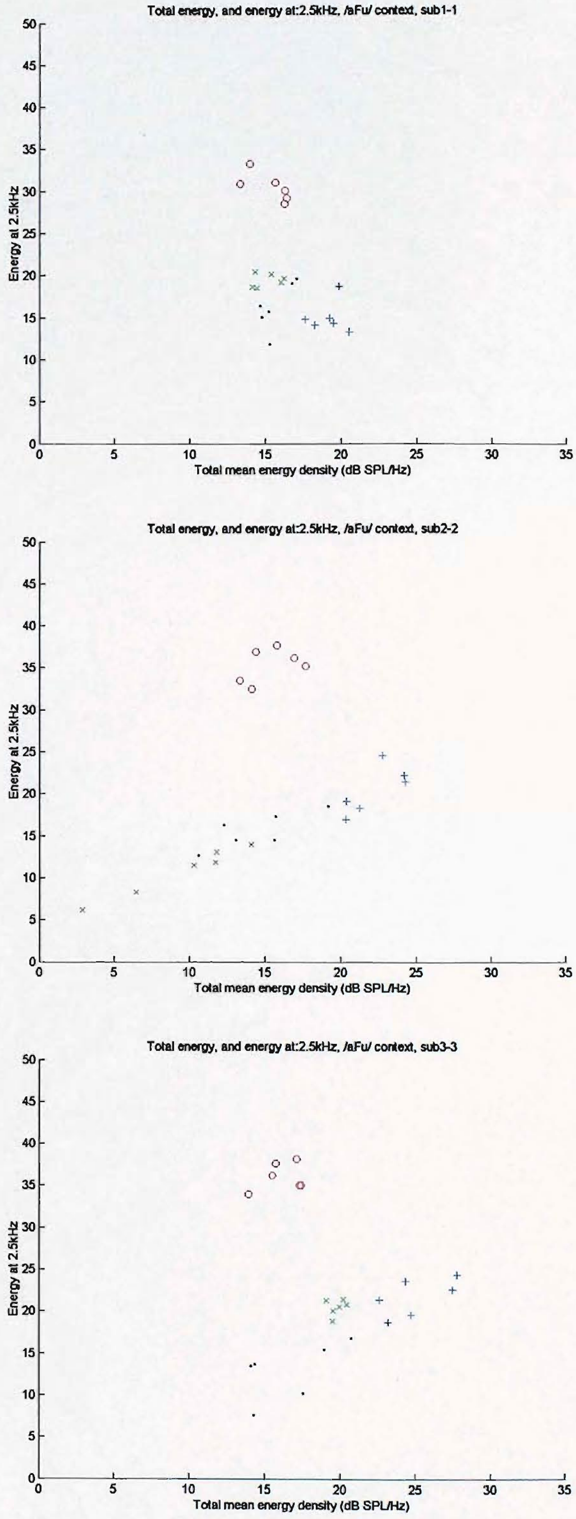


FIGURE F.11: Energy at 2.5 kHz against total spectral energy of all fricatives in /əFu/ context, subjects M-01 to M-03.

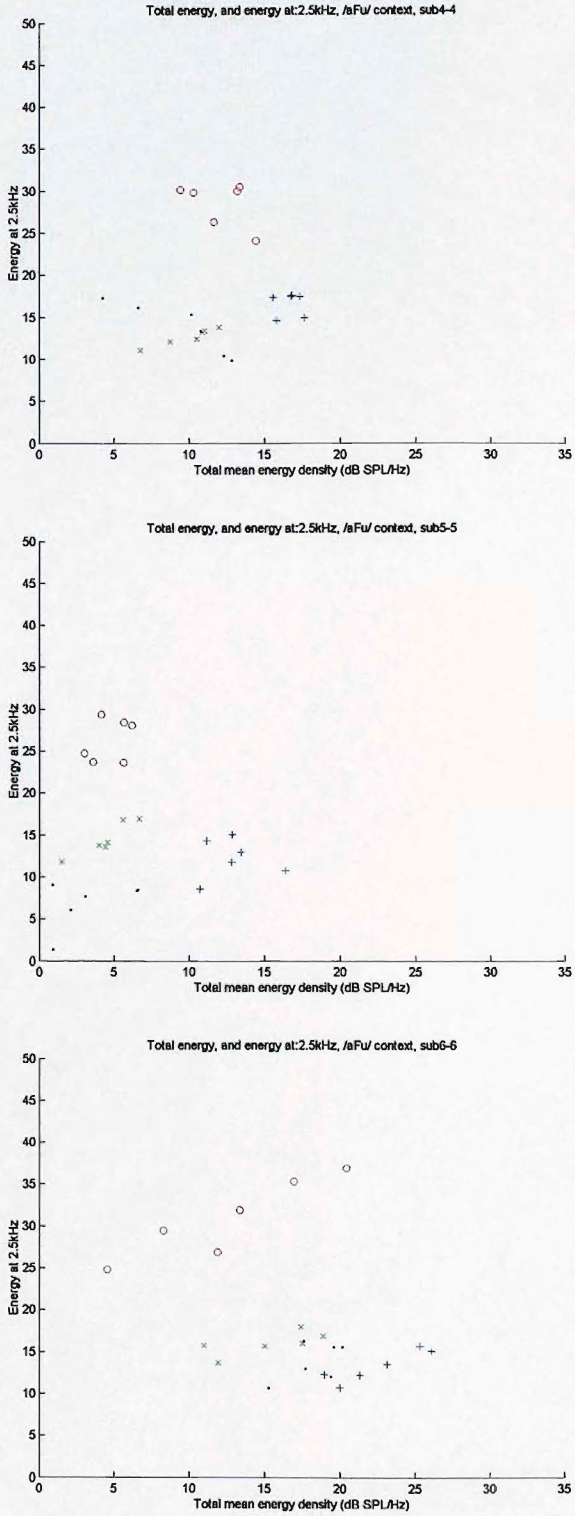


FIGURE F.12: Energy at 2.5 kHz against total spectral energy of all fricatives in /əFu/ context, subjects M-04 to M-06.

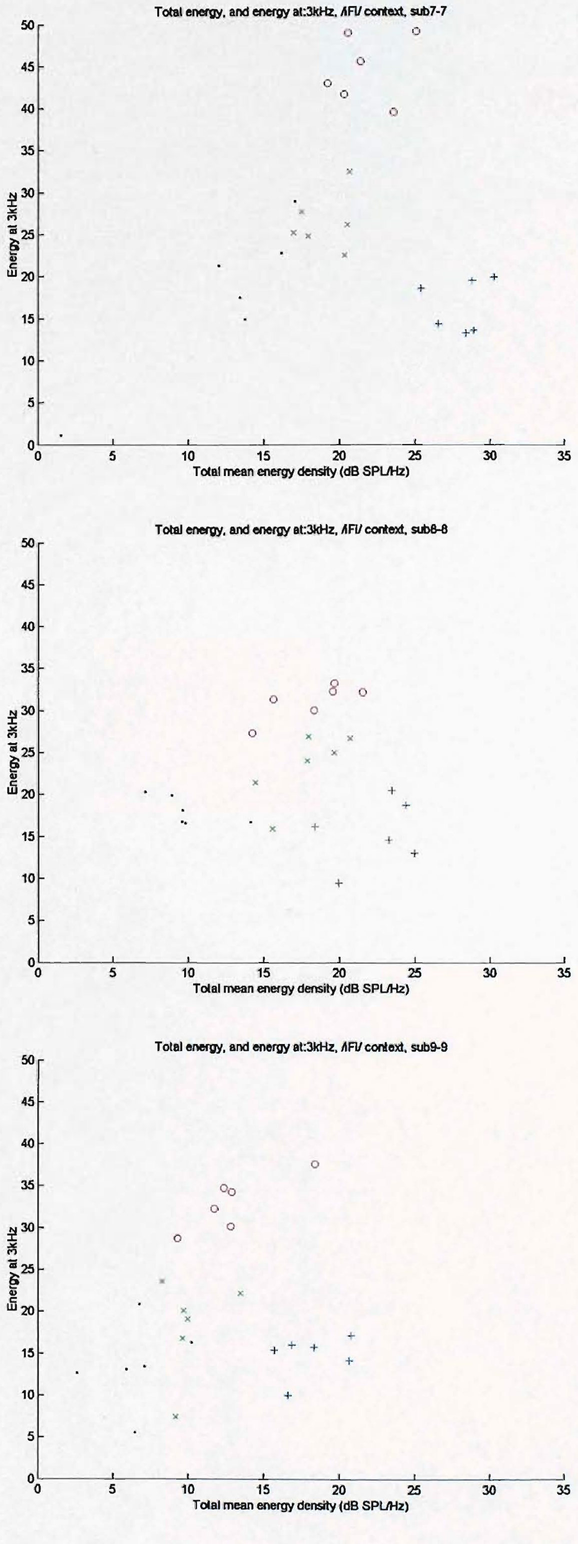


FIGURE F.13: Energy at 3 kHz against total spectral energy of all fricatives in /iFi/ context, subjects F-07 to F-09.

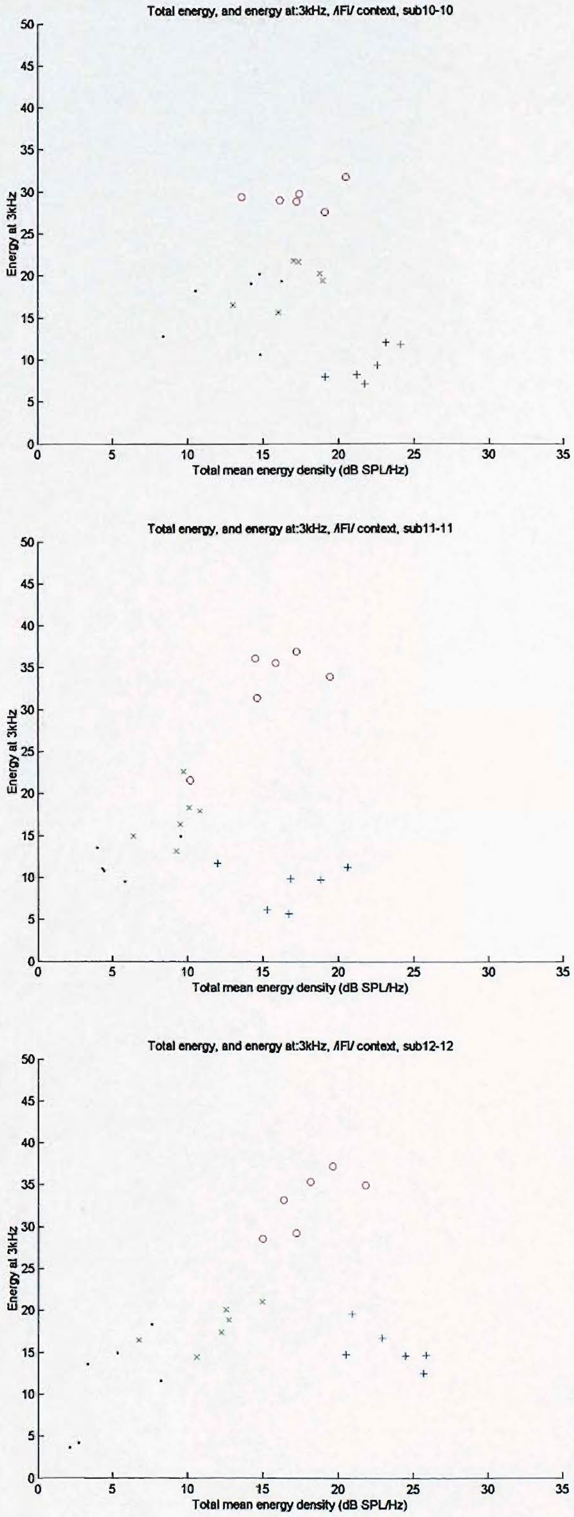


FIGURE F.14: Energy at 3 kHz against total spectral energy of all fricatives in /iFi/ context, subjects F-10 to F-12.

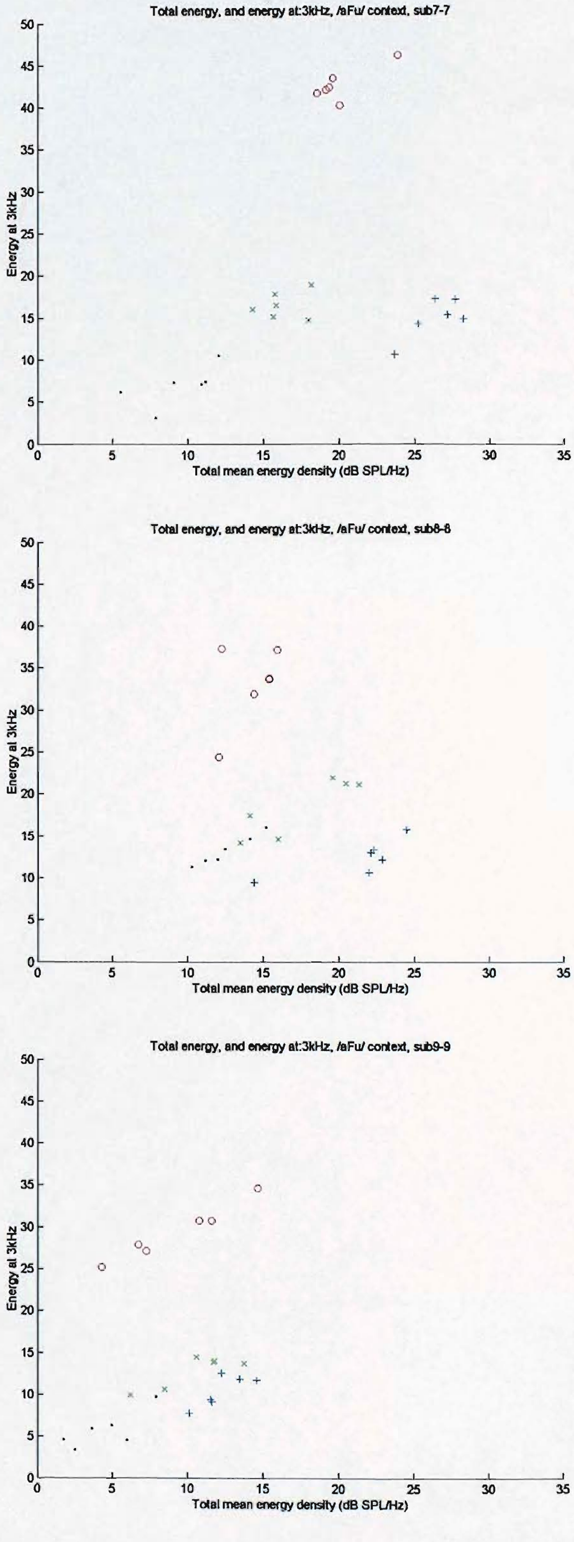


FIGURE F.15: Energy at 3 kHz against total spectral energy of all fricatives in /əFu/ context, subjects F-07 to F-09.

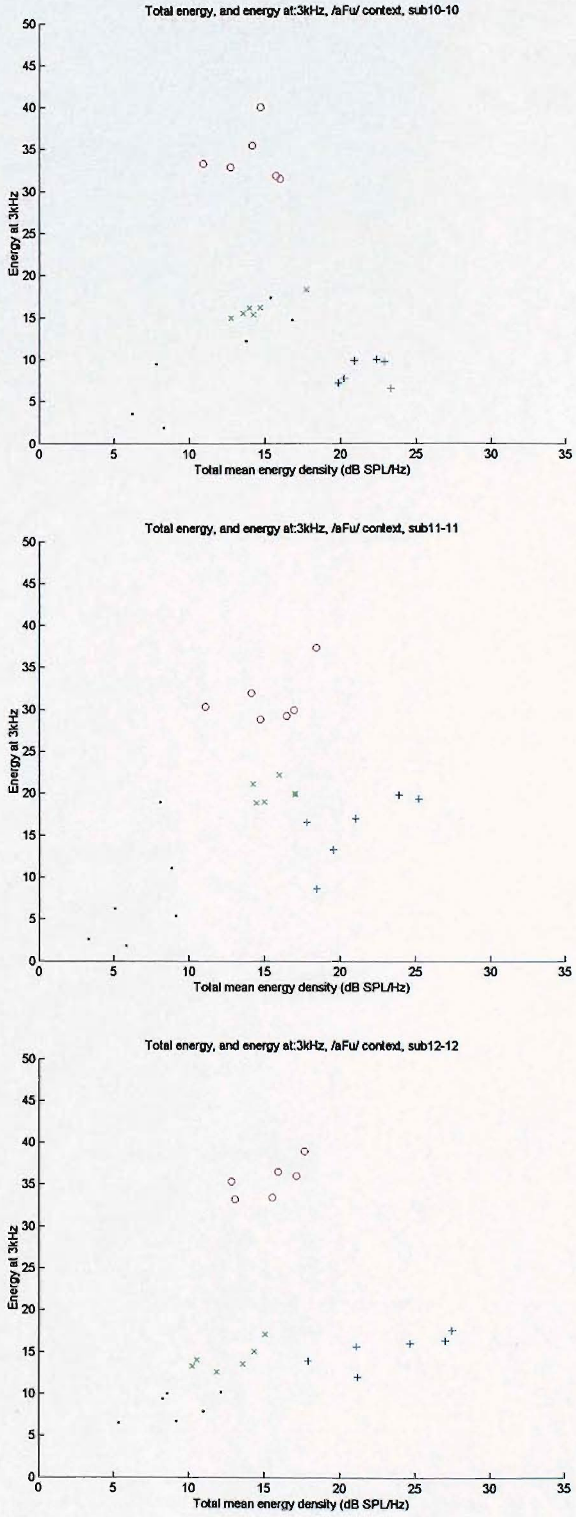


FIGURE F.16: Energy at 3 kHz against total spectral energy of all fricatives in /əFu/ context, subjects F-10 to F-12.

## F.3 Mean spectrum and spectral variability

All solid lines represent mean spectrum calculated over the full length of the token, while dashed lines show the spectral variance calculated over the full length of the token.

### F.3.1 Male example tokens

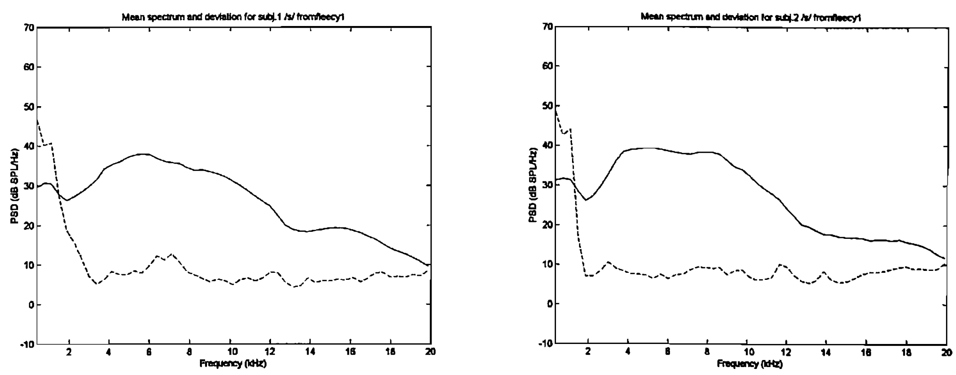


FIGURE F.17: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects M-01 and M-02.

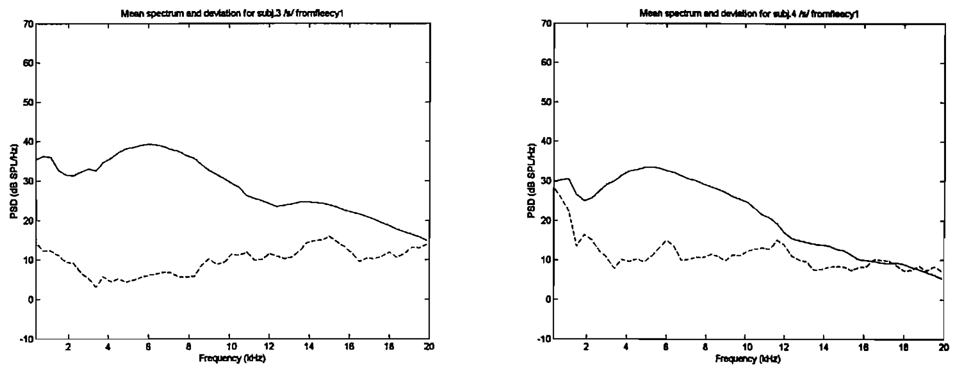


FIGURE F.18: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects M-03 and M-04.

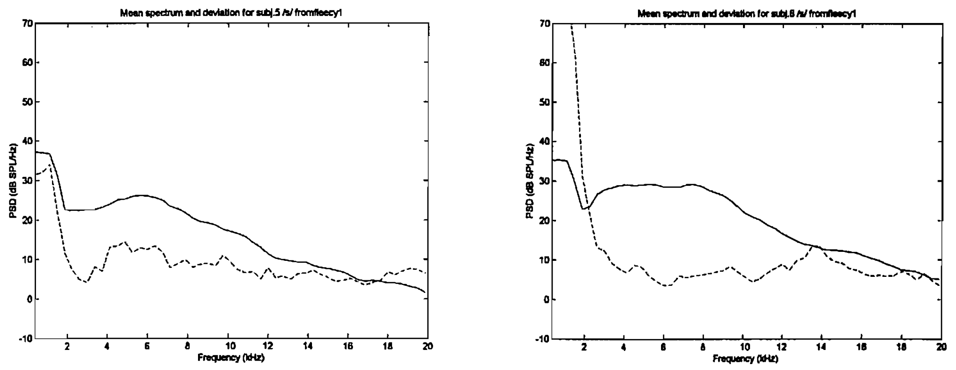


FIGURE F.19: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects M-05 and M-06.

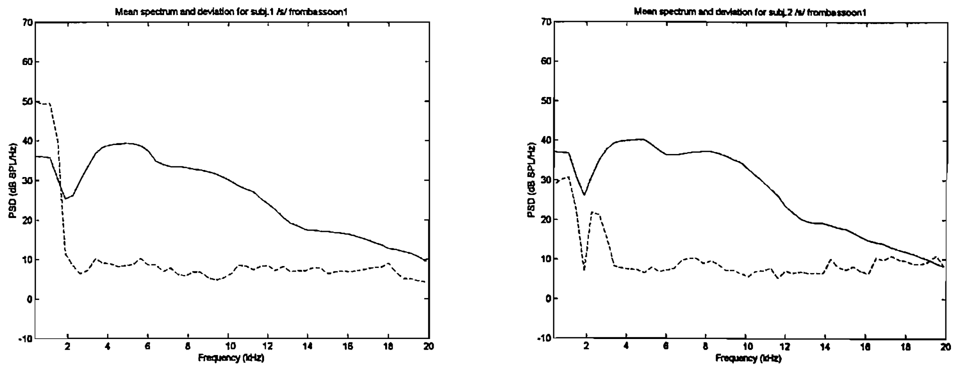


FIGURE F.20: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects M-01 and M-02.

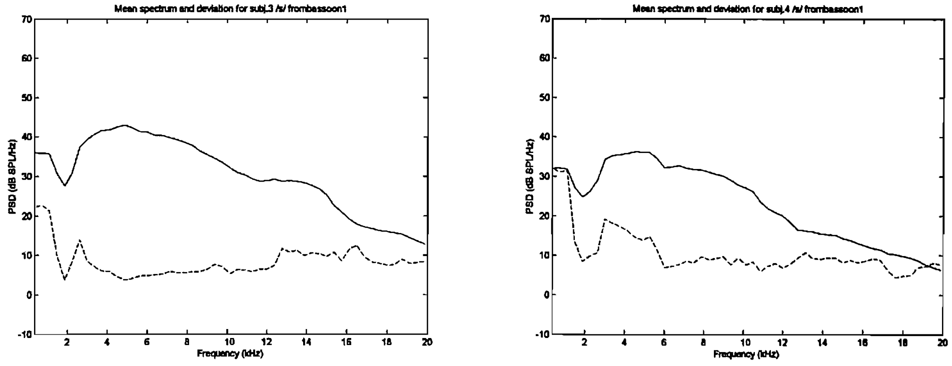


FIGURE F.21: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects M-03 and M-04.

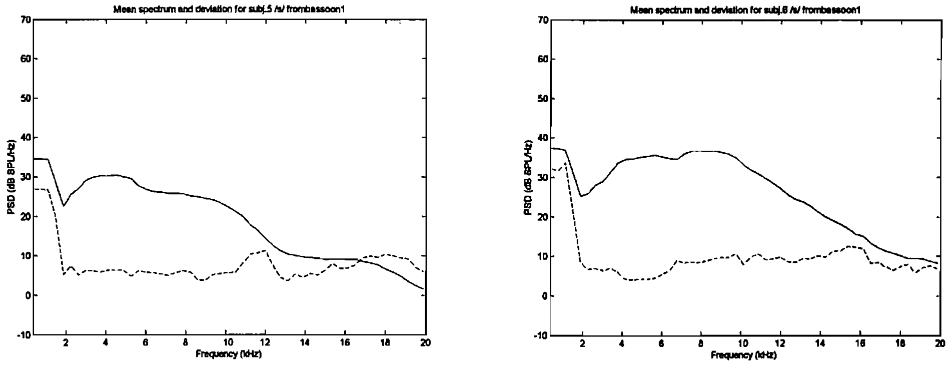


FIGURE F.22: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects M-05 and M-06.

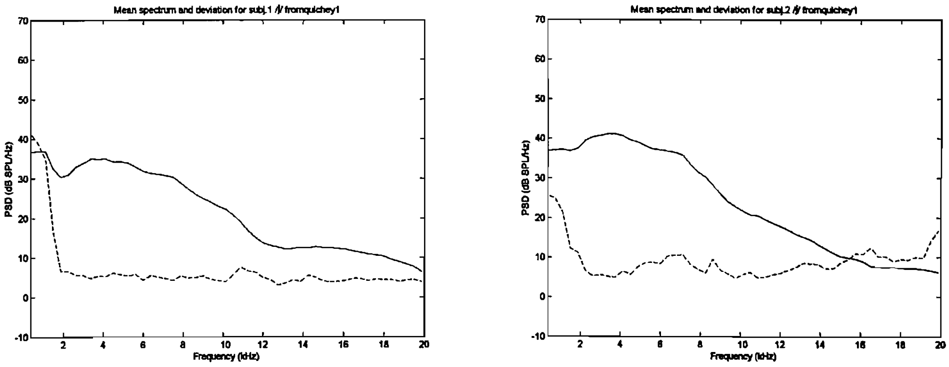


FIGURE F.23: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects M-01 and M-02.

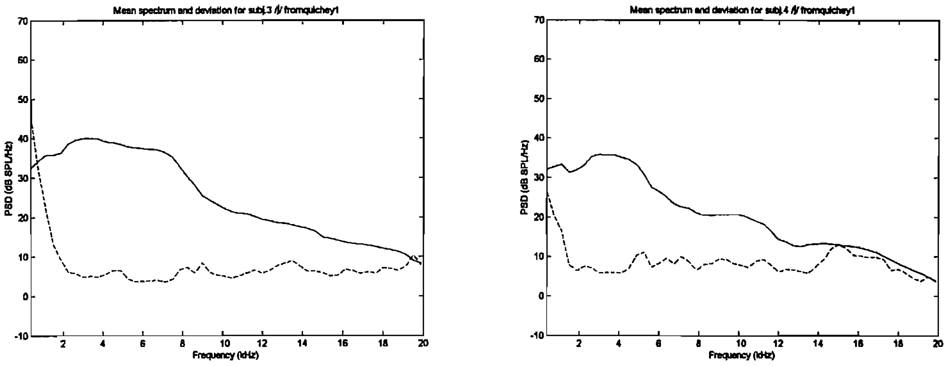


FIGURE F.24: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects M-03 and M-04.

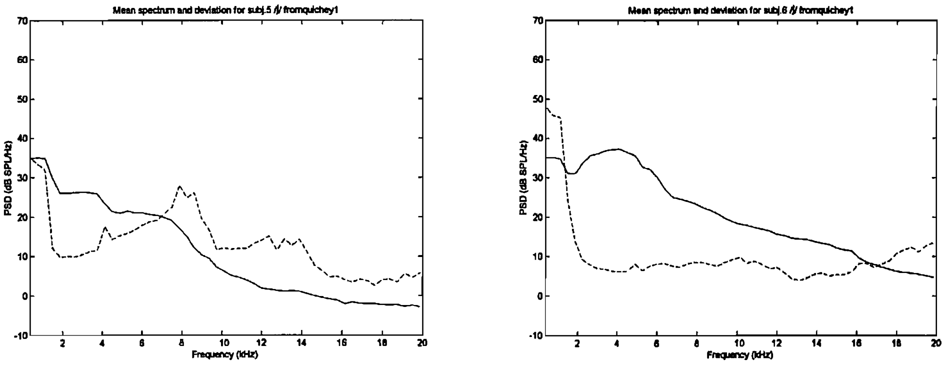


FIGURE F.25: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects M-05 and M-06.

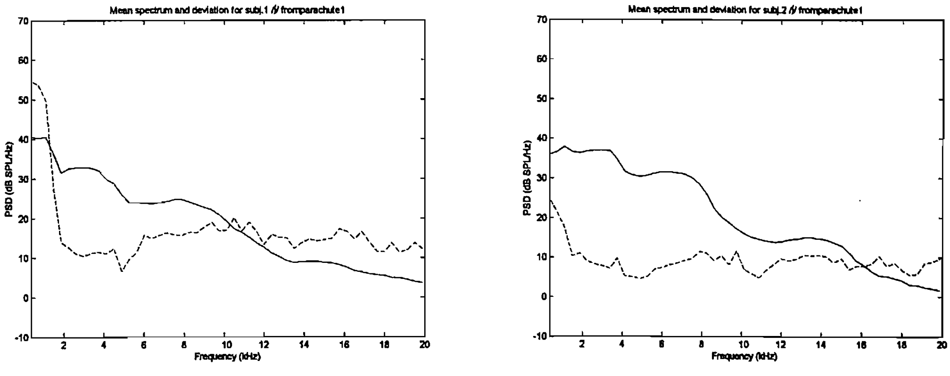


FIGURE F.26: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects M-01 and M-02.

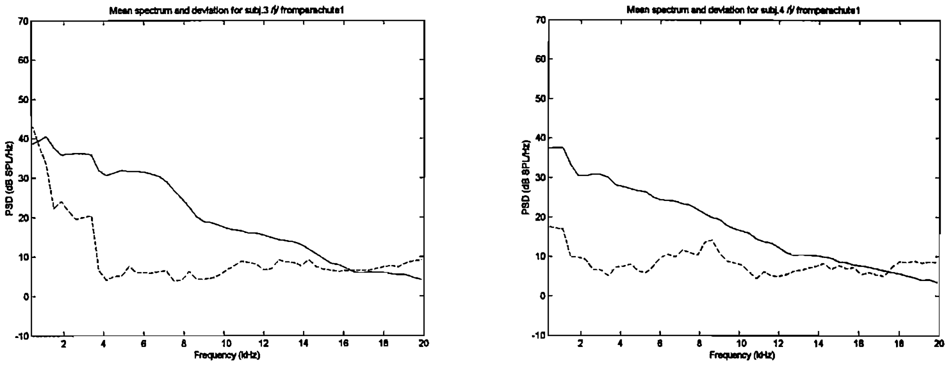


FIGURE F.27: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects M-03 and M-04.

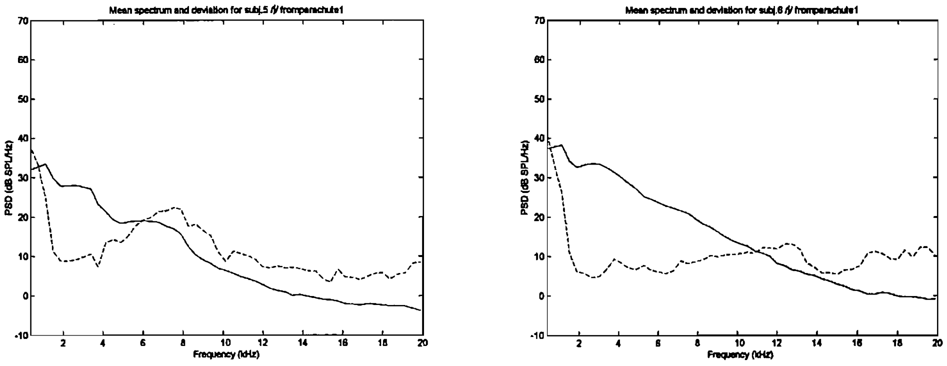


FIGURE F.28: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects M-05 and M-06.

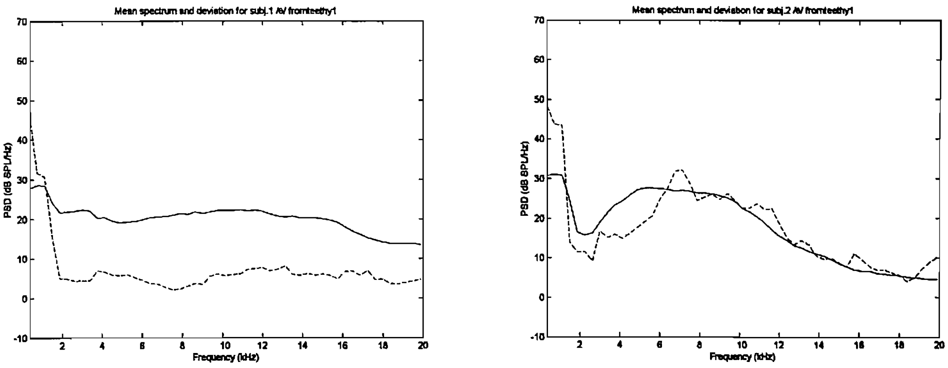


FIGURE F.29: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects M-01 and M-02.

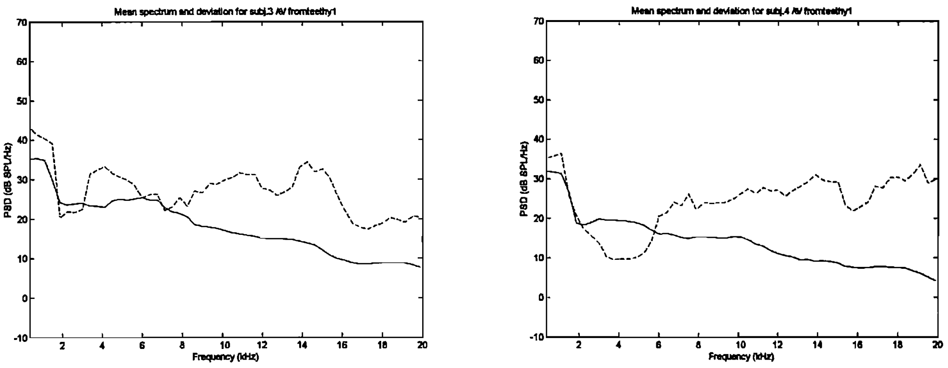


FIGURE F.30: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects M-03 and M-04.

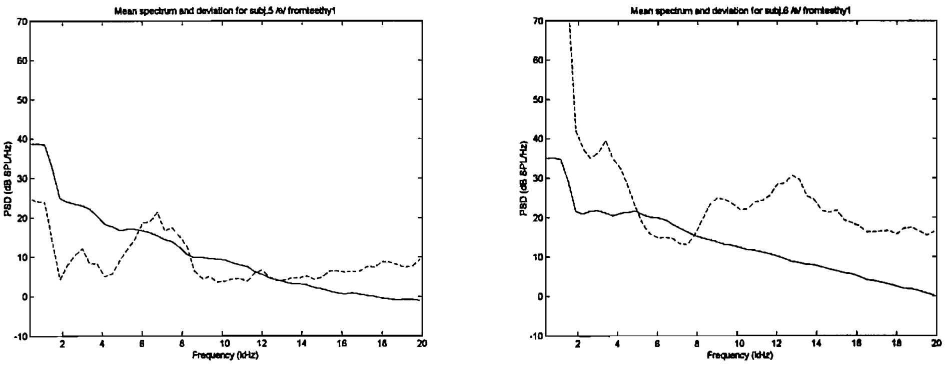


FIGURE F.31: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects M-05 and M-06.

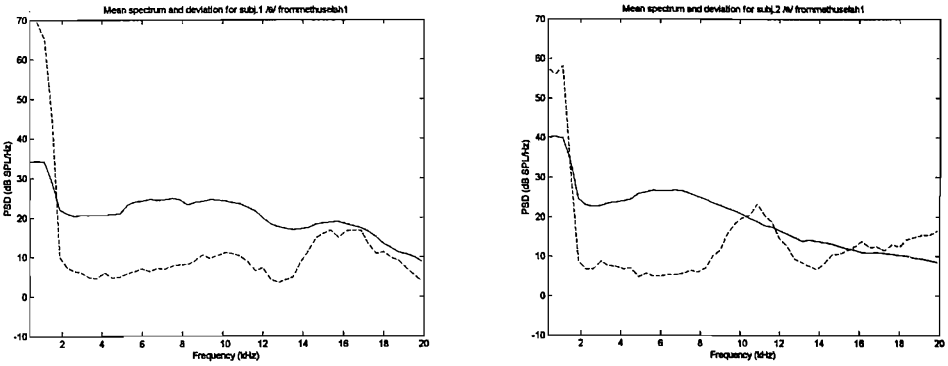


FIGURE F.32: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects M-01 and M-02.

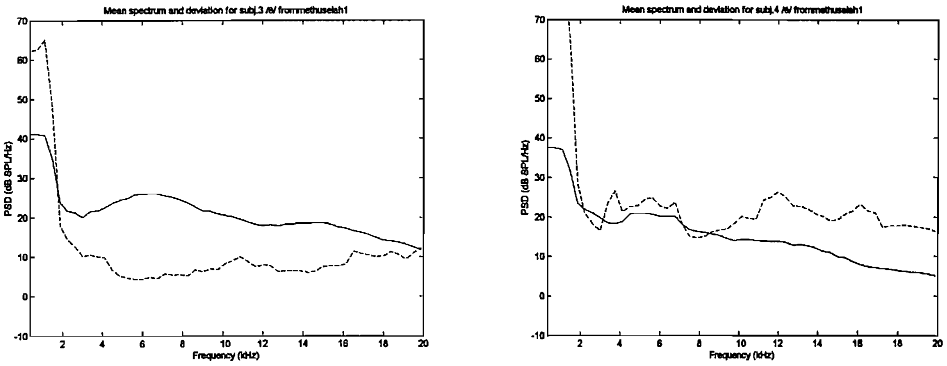


FIGURE F.33: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects M-03 and M-04.

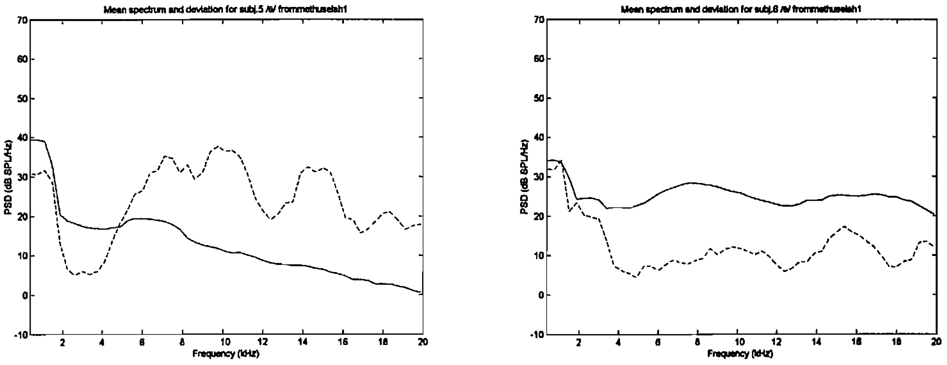


FIGURE F.34: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects M-05 and M-06.

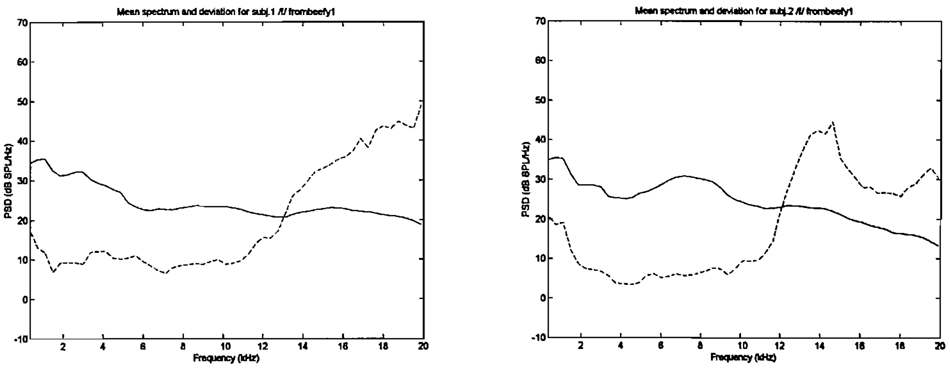


FIGURE F.35: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects M-01 and M-02.

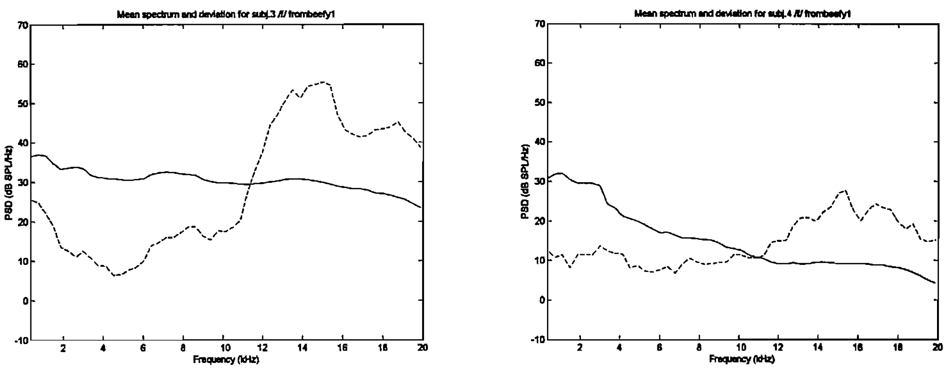


FIGURE F.36: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects M-03 and M-04.

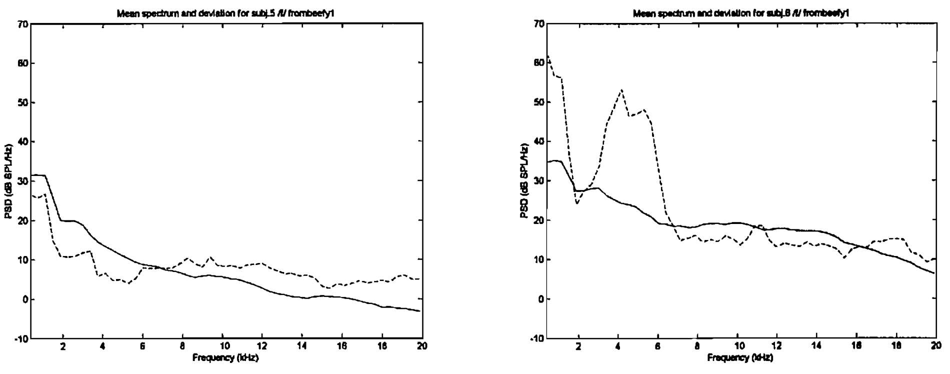


FIGURE F.37: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects M-05 and M-06.

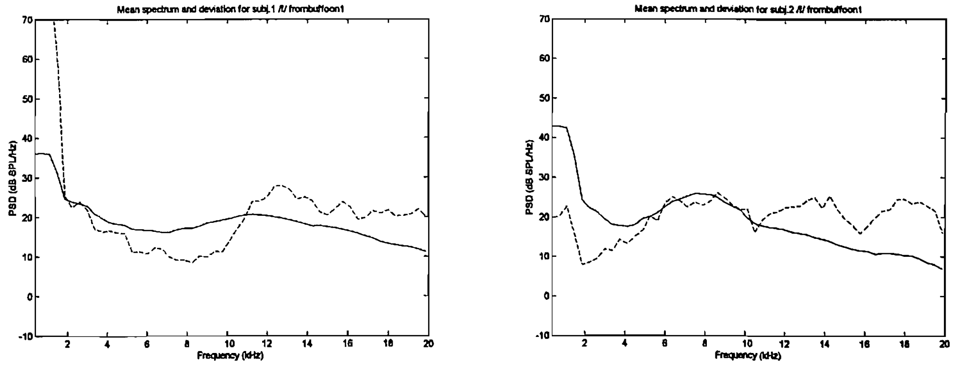


FIGURE F.38: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects M-01 and M-02.

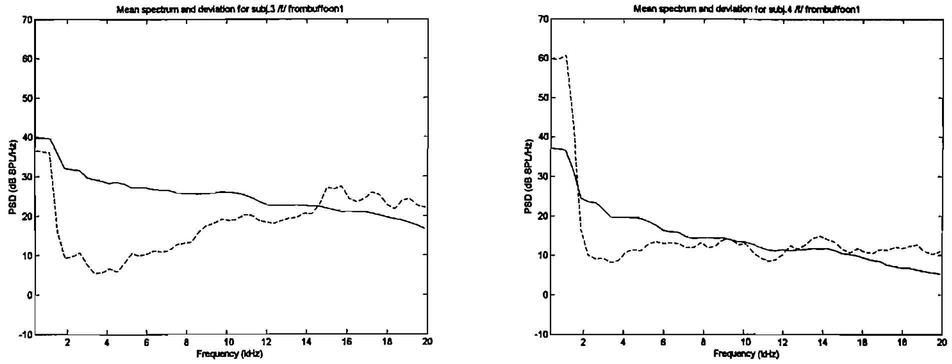


FIGURE F.39: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects M-03 and M-04.

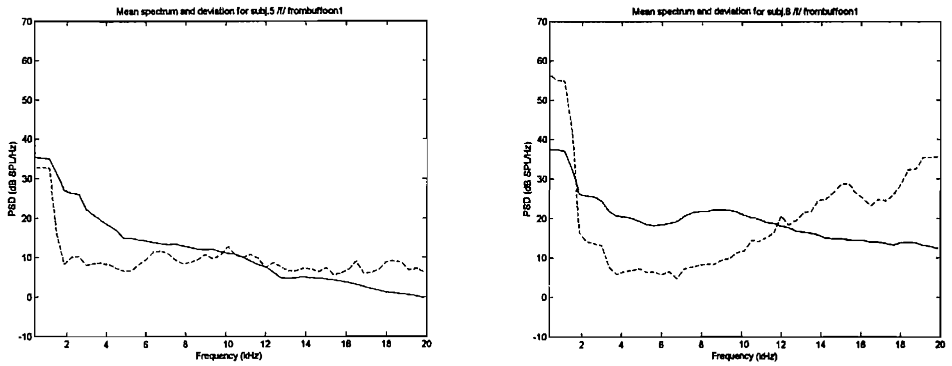


FIGURE F.40: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects M-05 and M-06.

### F.3.2 Female example tokens

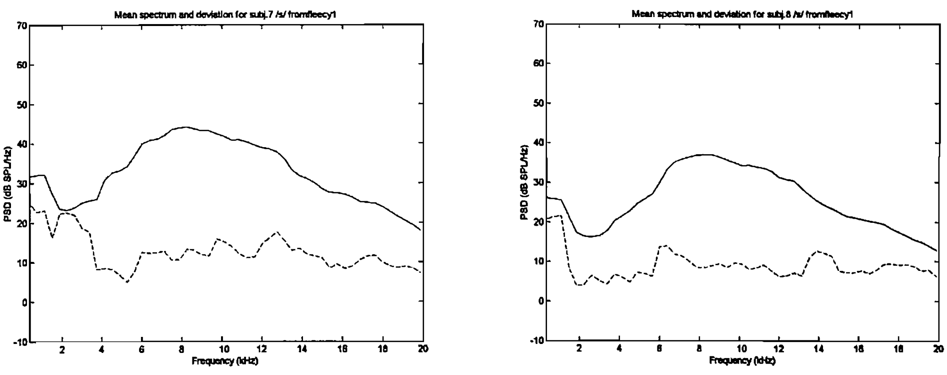


FIGURE F.41: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects F-07 and F-08.

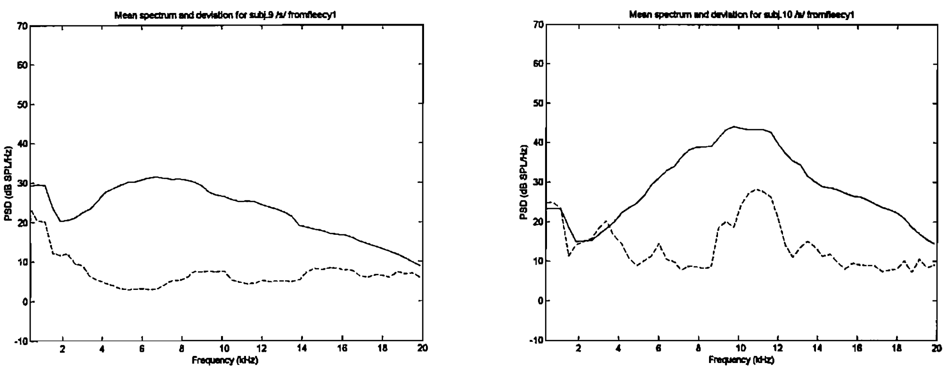


FIGURE F.42: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects F-09 and F-10.

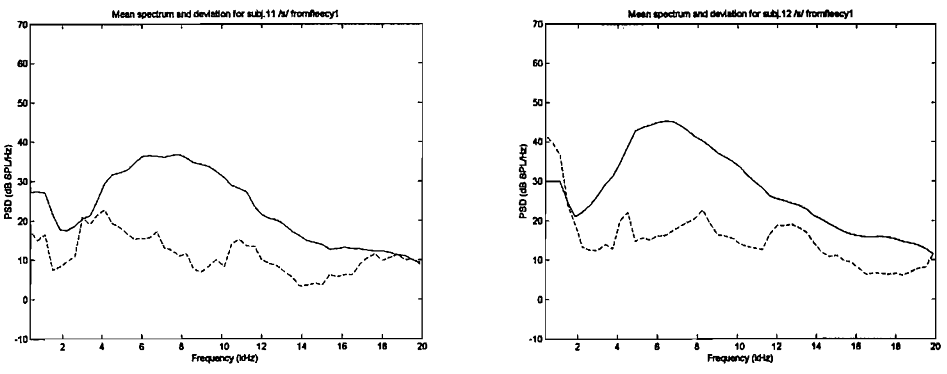


FIGURE F.43: Mean spectrum and spectral variability calculated over time of example /isi/ tokens, subjects F-11 and F-12.

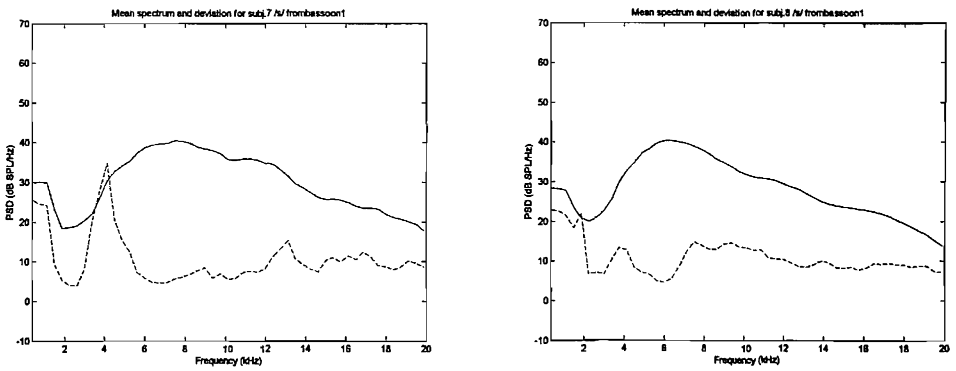


FIGURE F.44: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects F-07 and F-08.

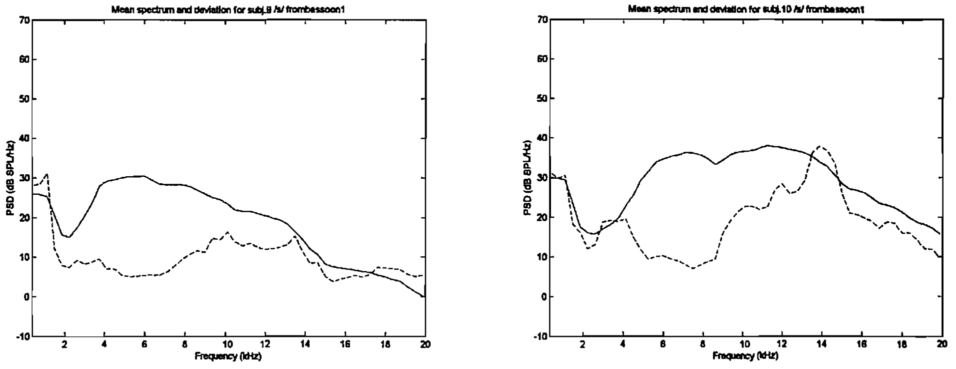


FIGURE F.45: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects F-09 and F-10.

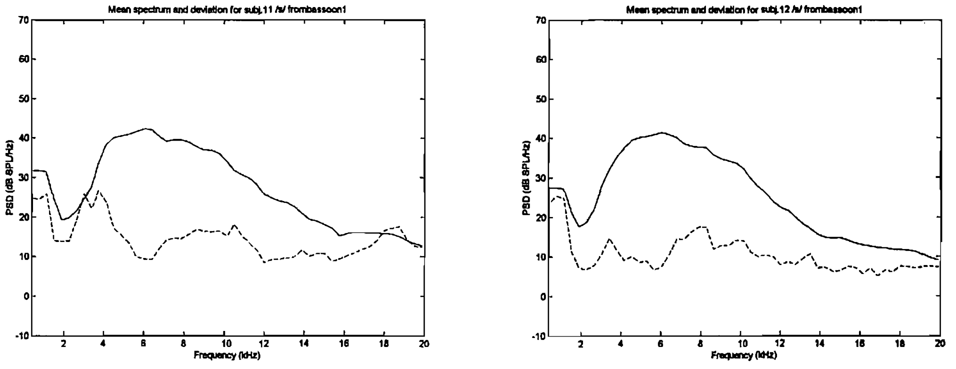


FIGURE F.46: Mean spectrum and spectral variability calculated over time of example /əsu/ tokens, subjects F-11 and F-12.

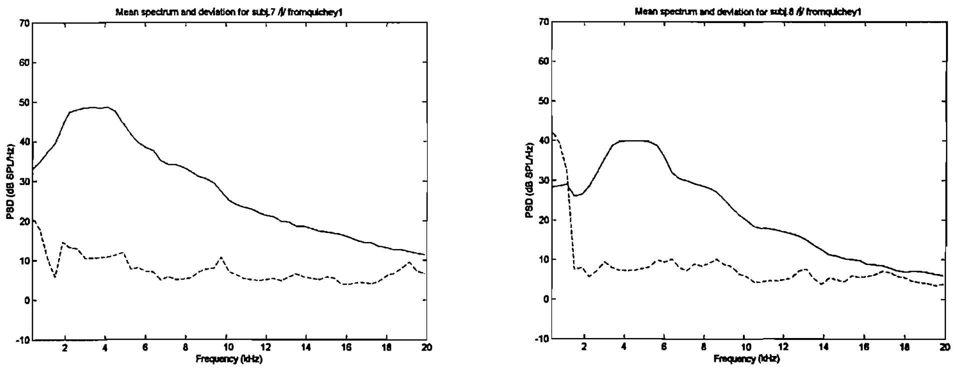


FIGURE F.47: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects F-07 and F-08.

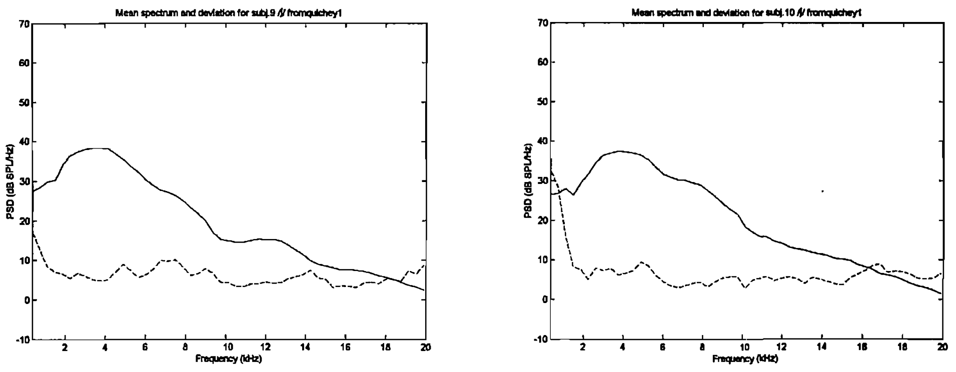


FIGURE F.48: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects F-09 and F-10.

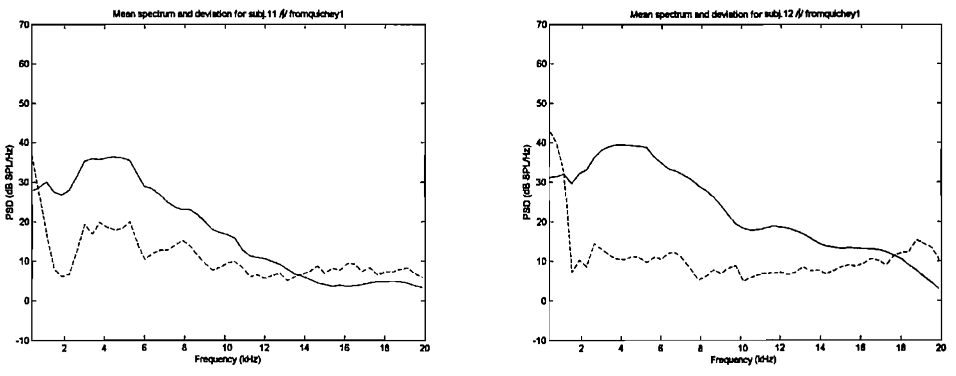


FIGURE F.49: Mean spectrum and spectral variability calculated over time of example /iʃi/ tokens, subjects F-11 and F-12.

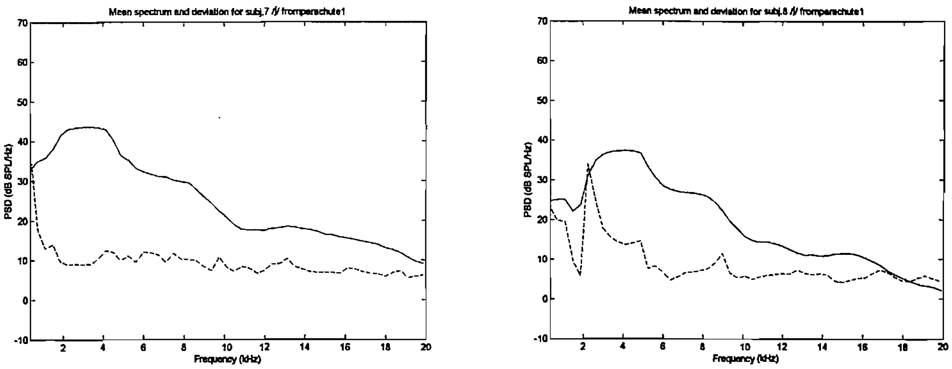


FIGURE F.50: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects F-07 and F-08.

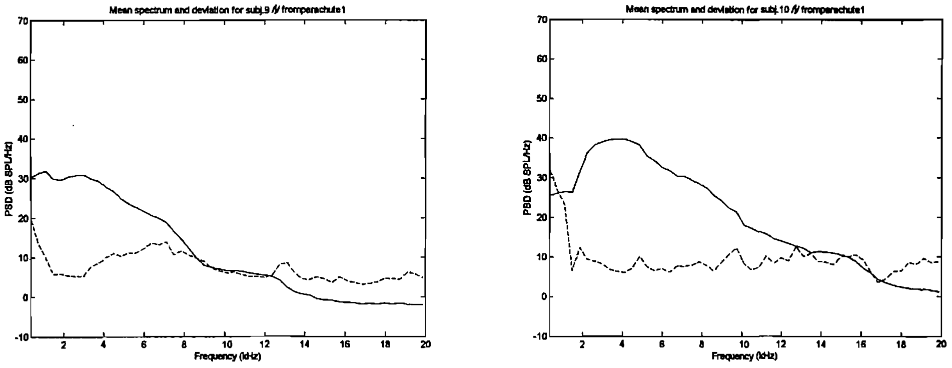


FIGURE F.51: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects F-09 and F-10.

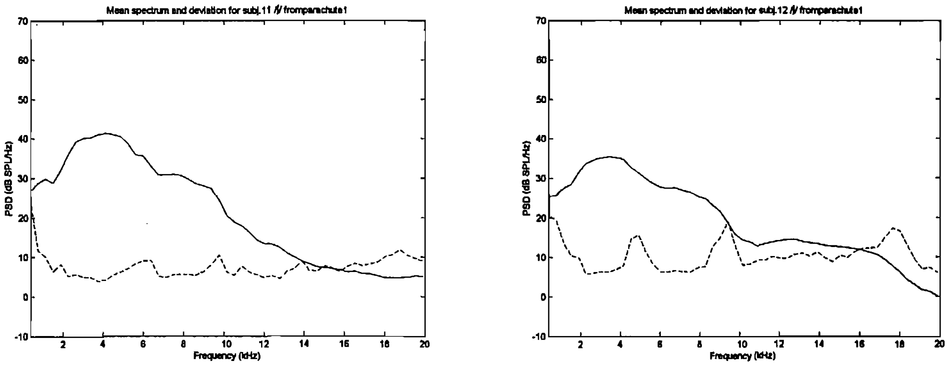


FIGURE F.52: Mean spectrum and spectral variability calculated over time of example /əʃu/ tokens, subjects F-11 and F-12.

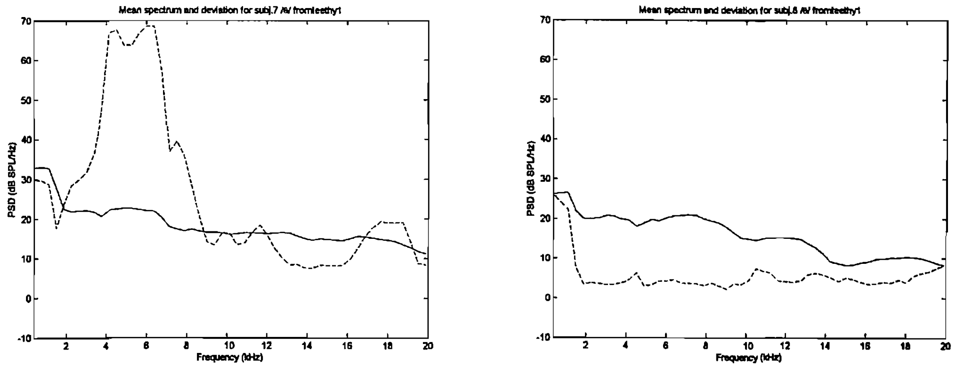


FIGURE F.53: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects F-07 and F-08.

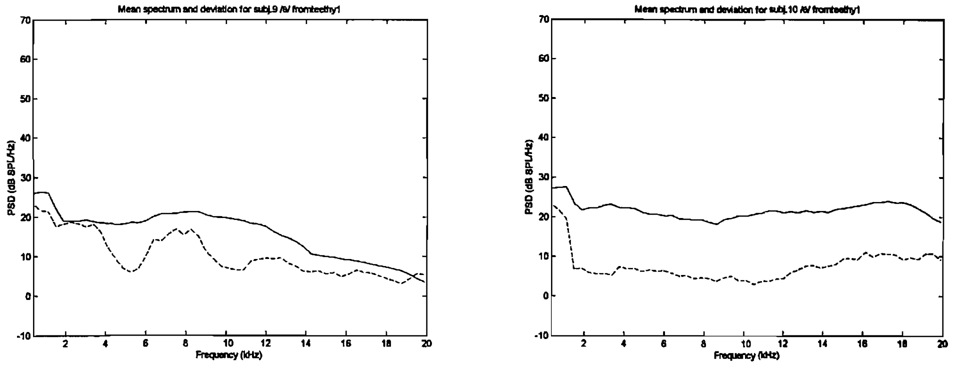


FIGURE F.54: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects F-09 and F-10.

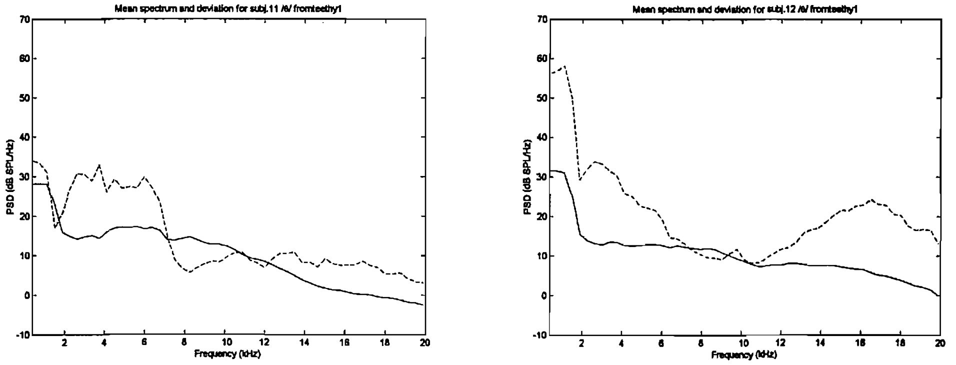


FIGURE F.55: Mean spectrum and spectral variability calculated over time of example /iθi/ tokens, subjects F-11 and F-12.

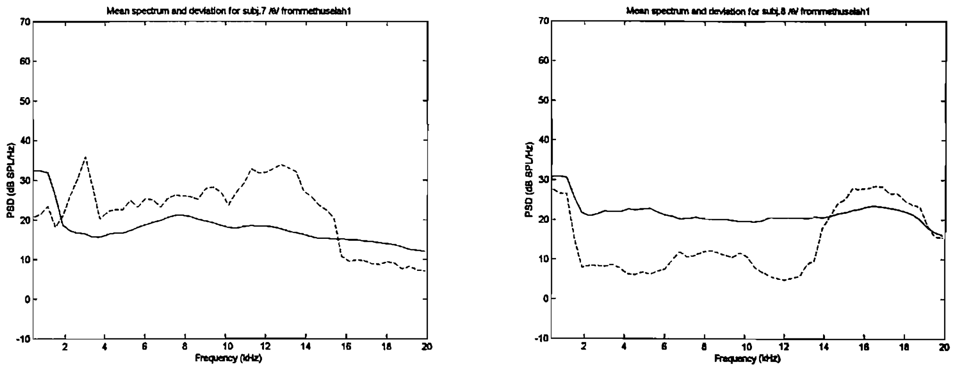


FIGURE F.56: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects F-07 and F-08.

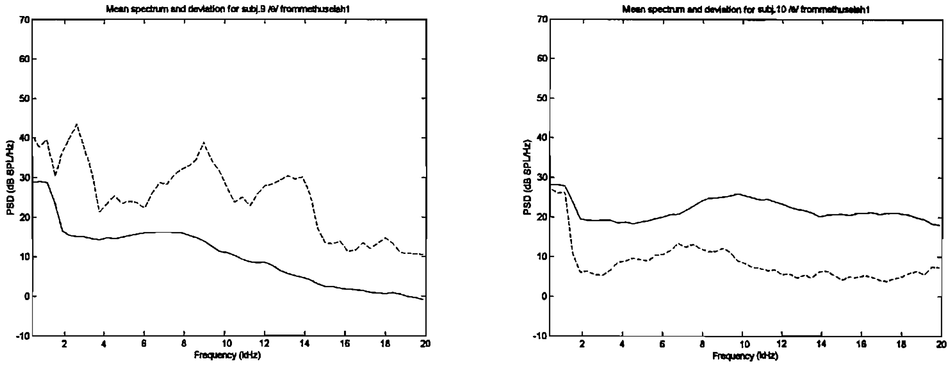


FIGURE F.57: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects F-09 and F-10.

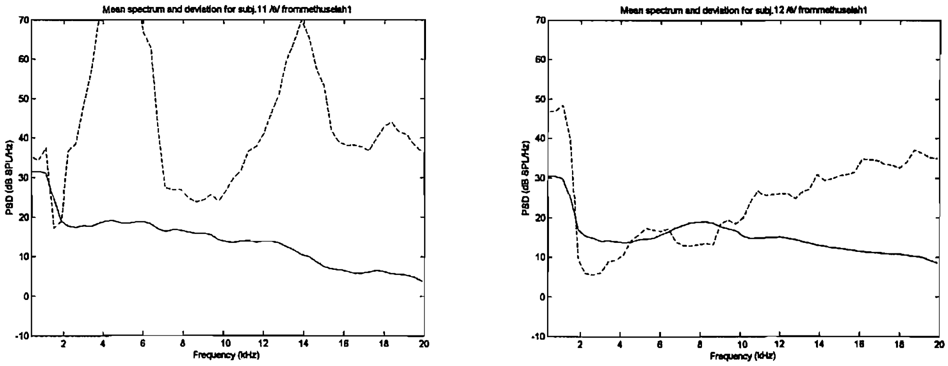


FIGURE F.58: Mean spectrum and spectral variability calculated over time of example /əθu/ tokens, subjects F-11 and F-12.

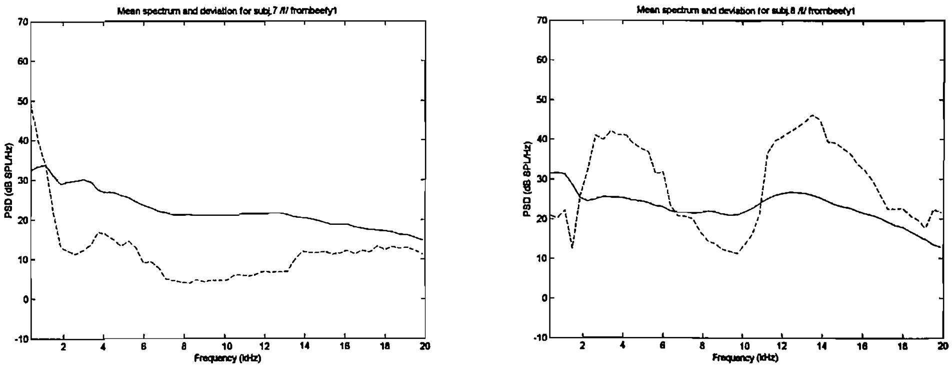


FIGURE F.59: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects F-07 and F-08.

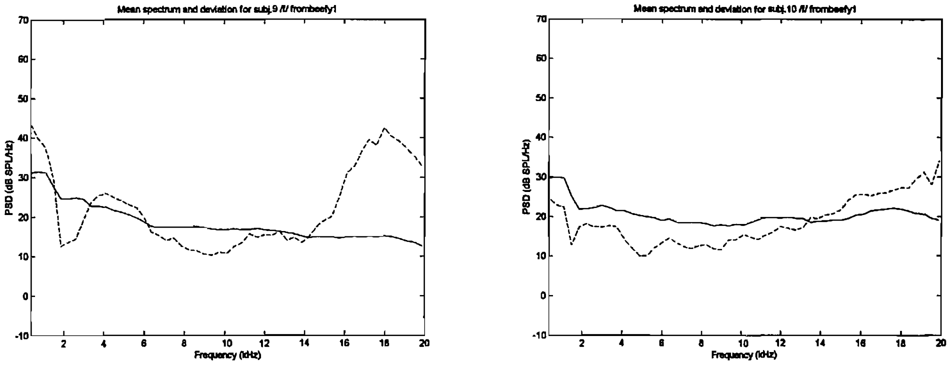


FIGURE F.60: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects F-09 and F-10.

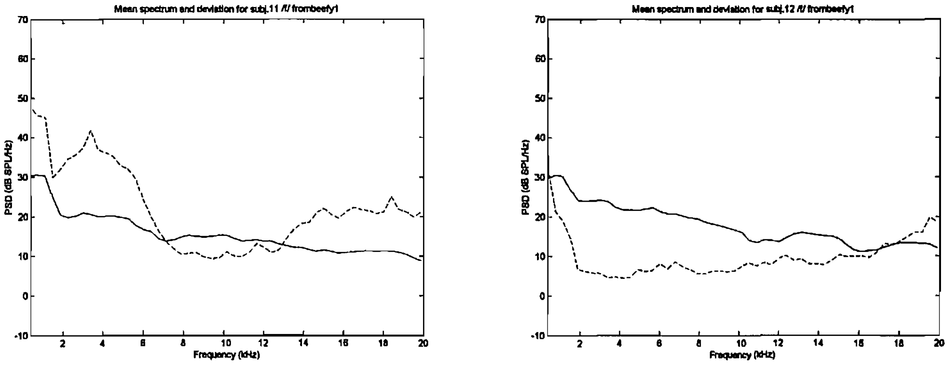


FIGURE F.61: Mean spectrum and spectral variability calculated over time of example /ifi/ tokens, subjects F-11 and F-12.

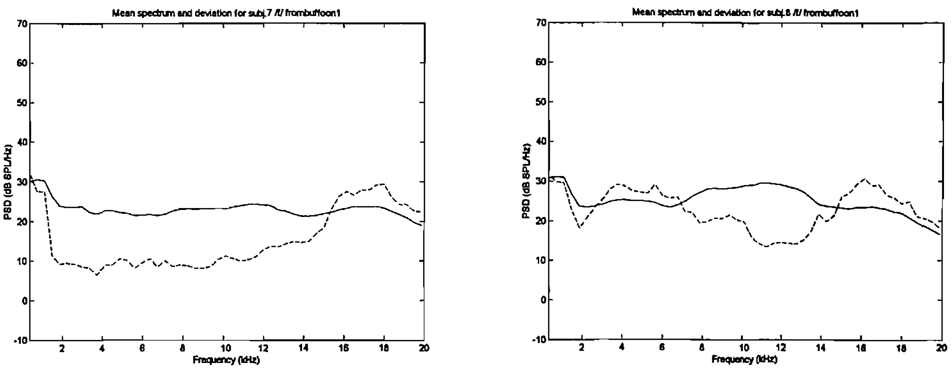


FIGURE F.62: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects F-07 and F-08.

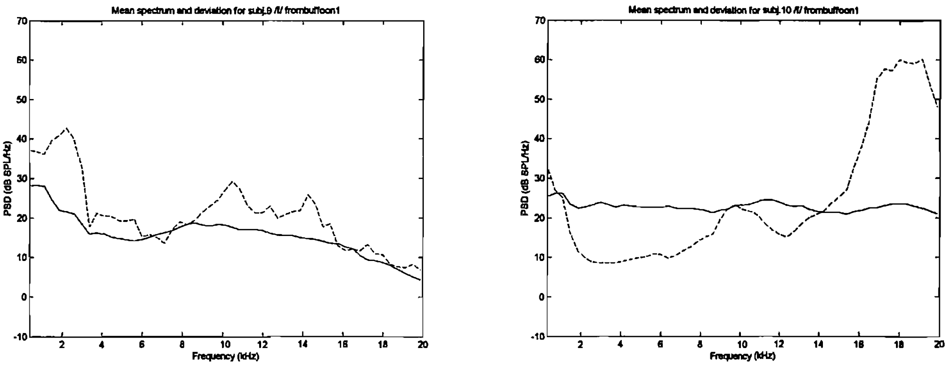


FIGURE F.63: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects F-09 and F-10.

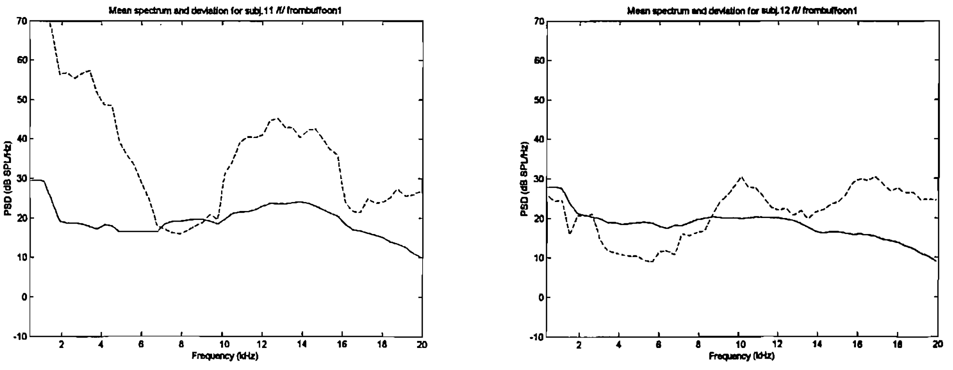


FIGURE F.64: Mean spectrum and spectral variability calculated over time of example /əfu/ tokens, subjects F-11 and F-12.

## F.4 Total spectral variability by subject and example vowel context

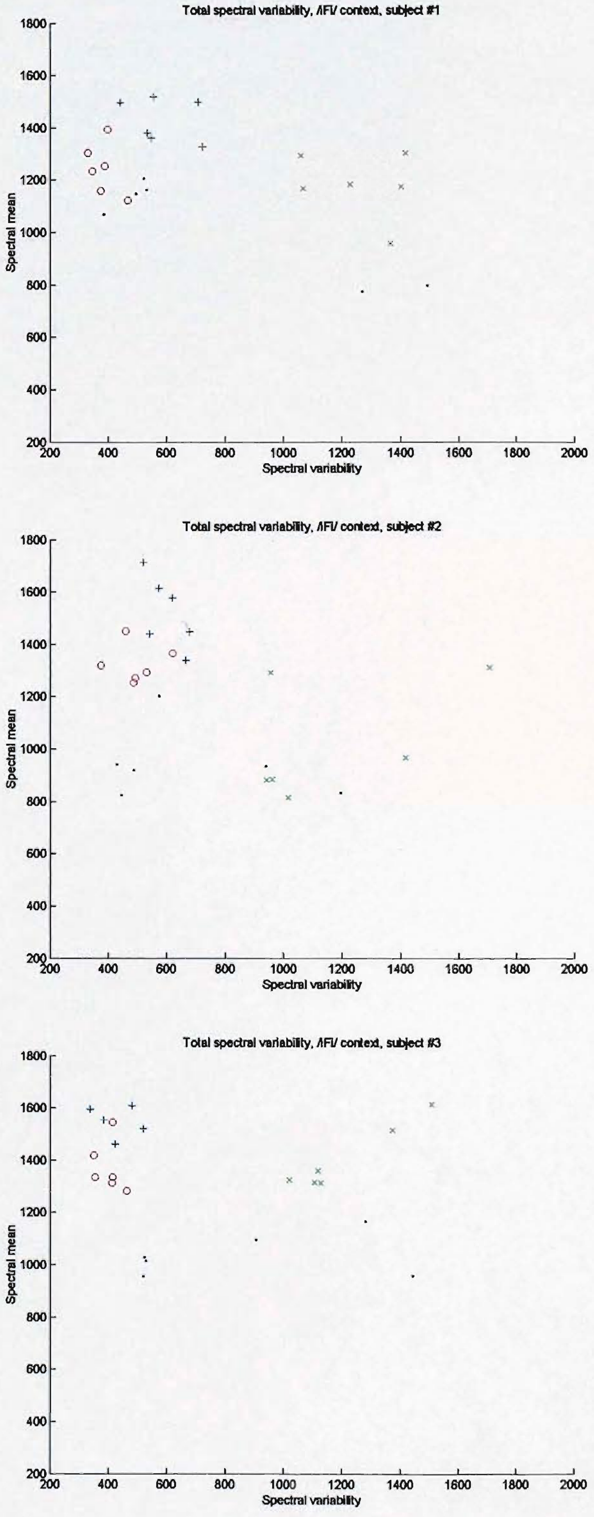


FIGURE F.65: Total spectral variability of all fricatives in /iFi/ context, subjects M-01 to M-03.

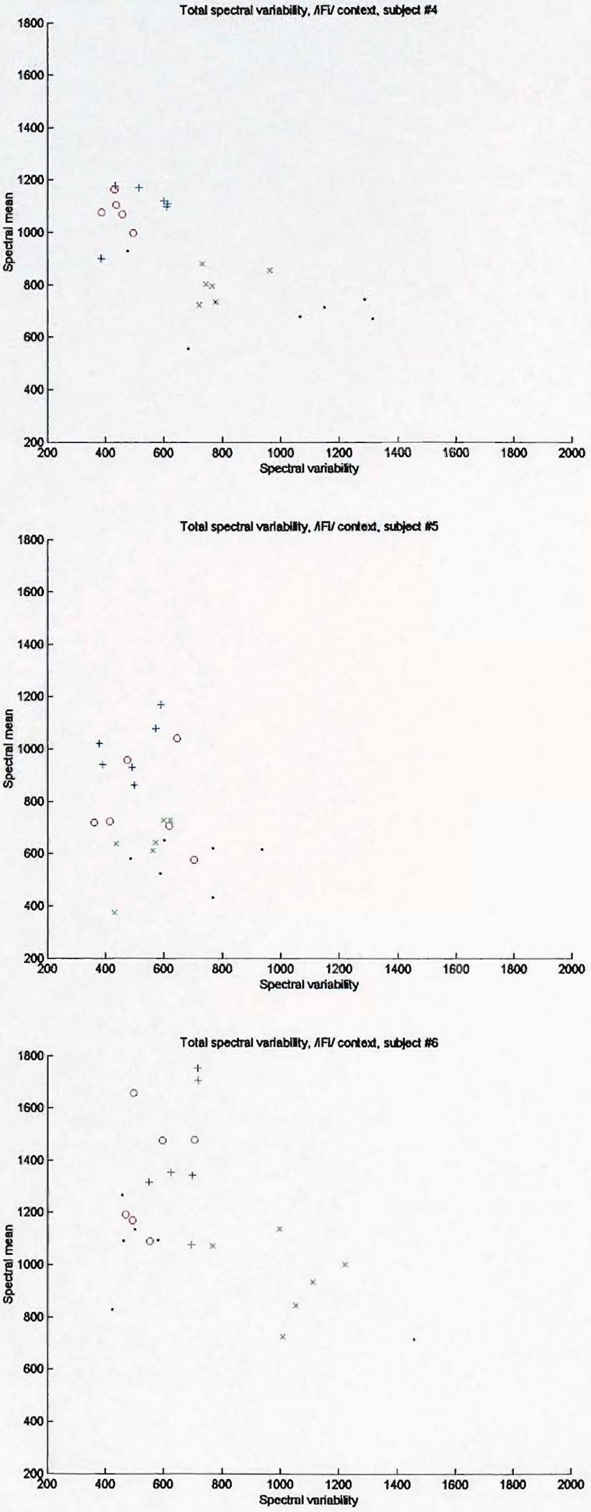


FIGURE F.66: Total spectral variability of all fricatives in /iFi/ context, subjects M-04 to M-06.

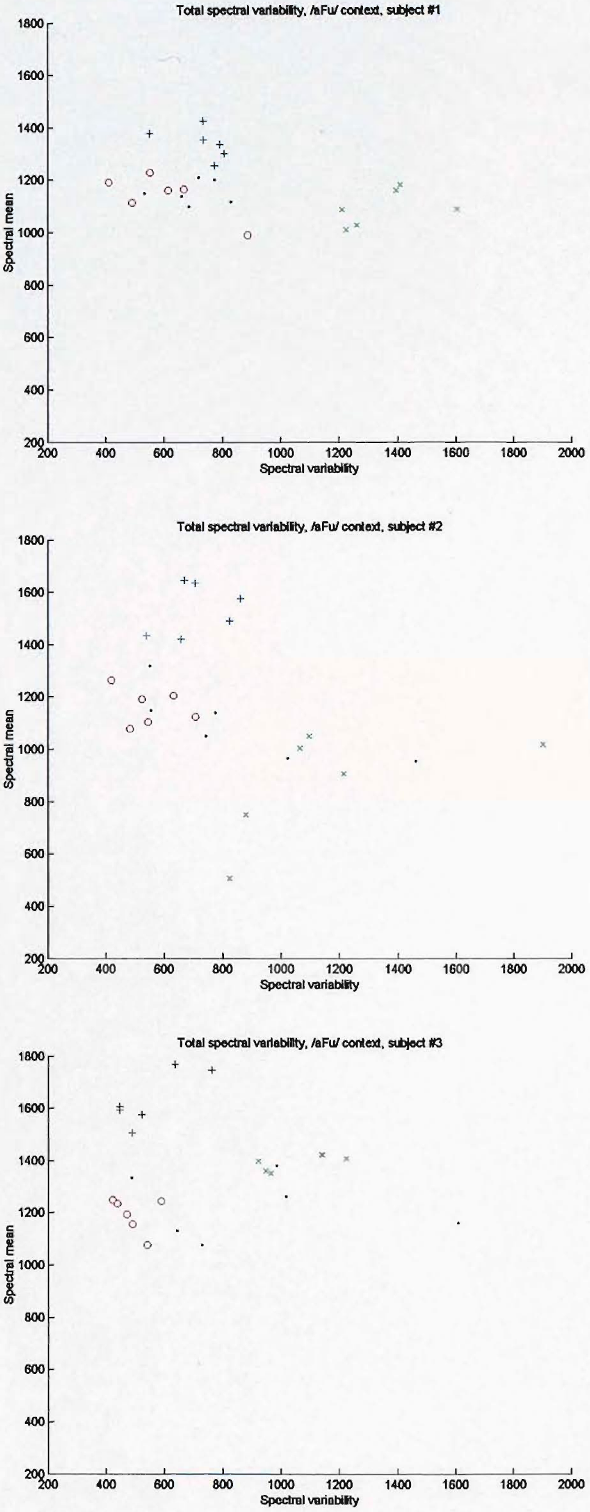


FIGURE F.67: Total spectral variability of all fricatives in /əFu/ context, subjects M-01 to M-03.

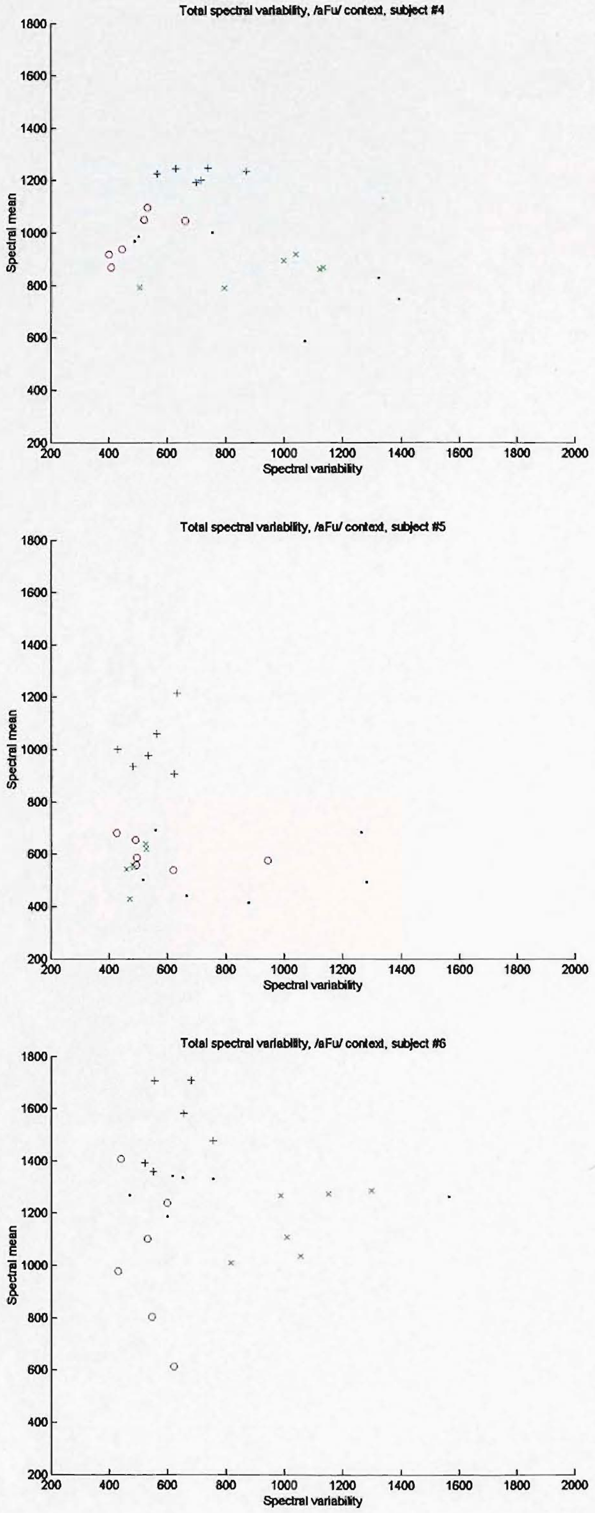


FIGURE F.68: Total spectral variability of all fricatives in /əFu/ context, subjects M-04 to M-06.

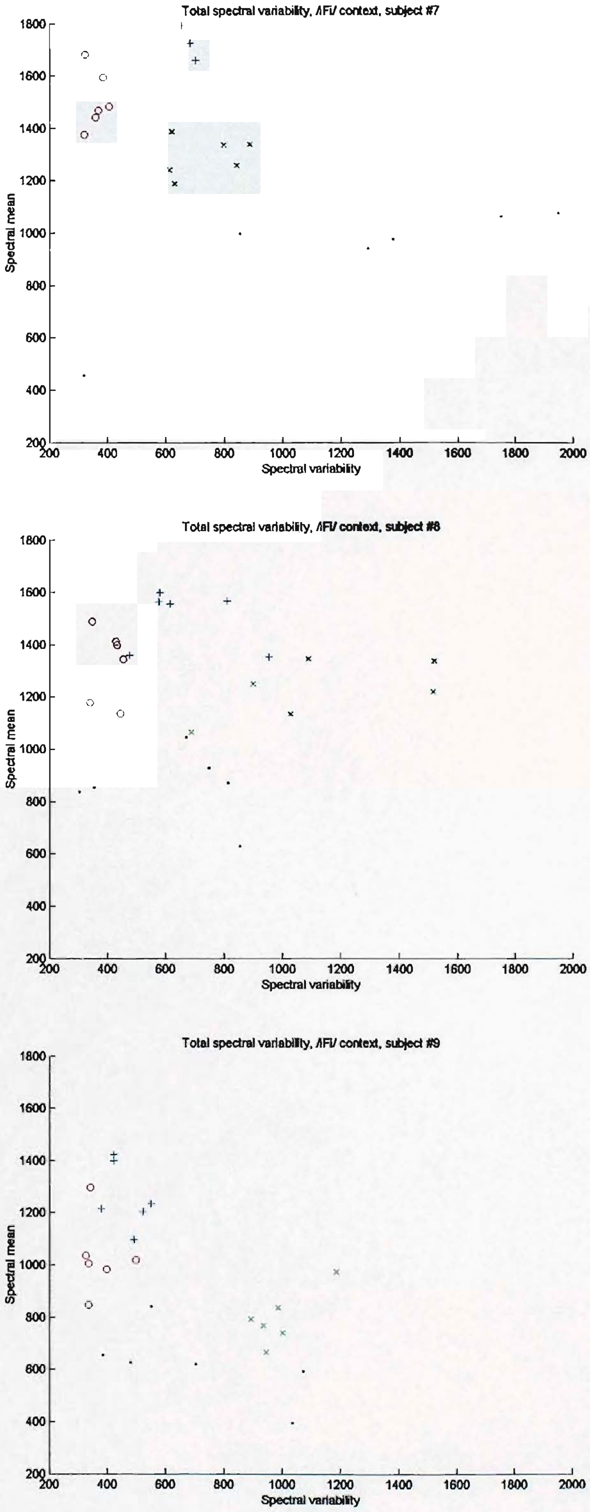


FIGURE F.69: Total spectral variability of all fricatives in /iFi/ context, subjects F-07 to F-09.

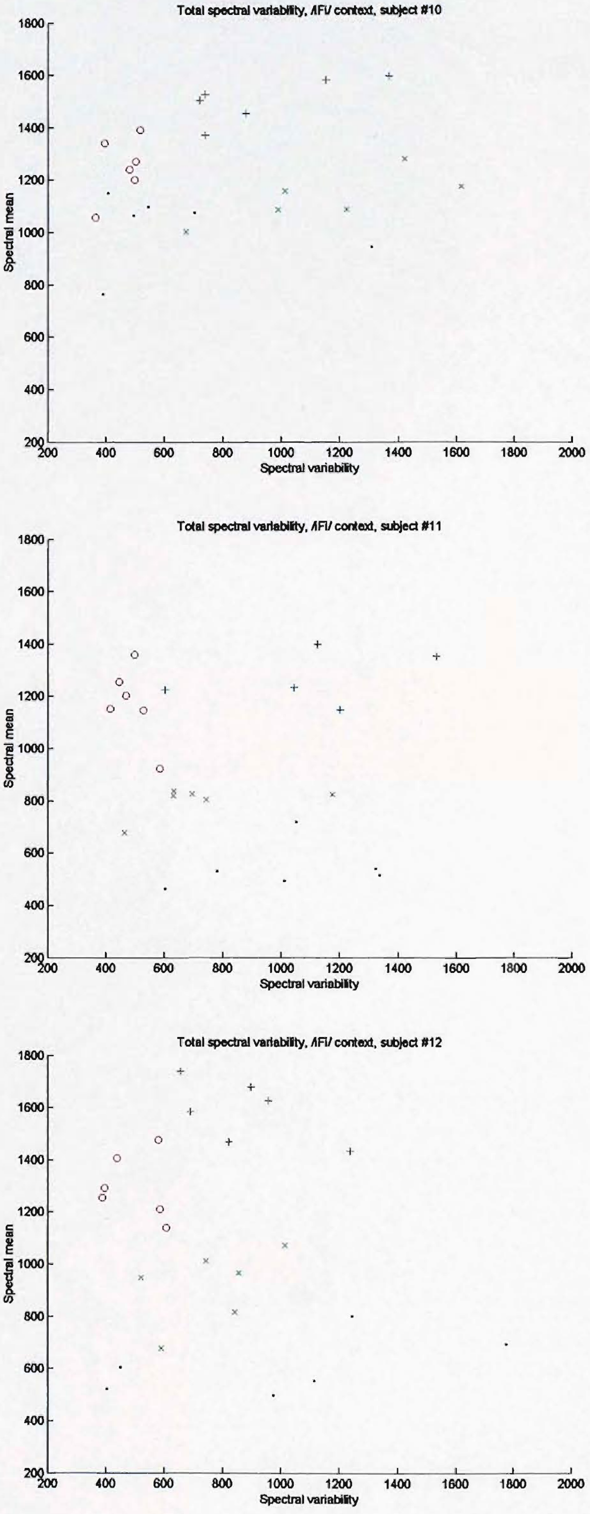


FIGURE F.70: Total spectral variability of all fricatives in /iFi/ context, subjects F-10 to F-12.

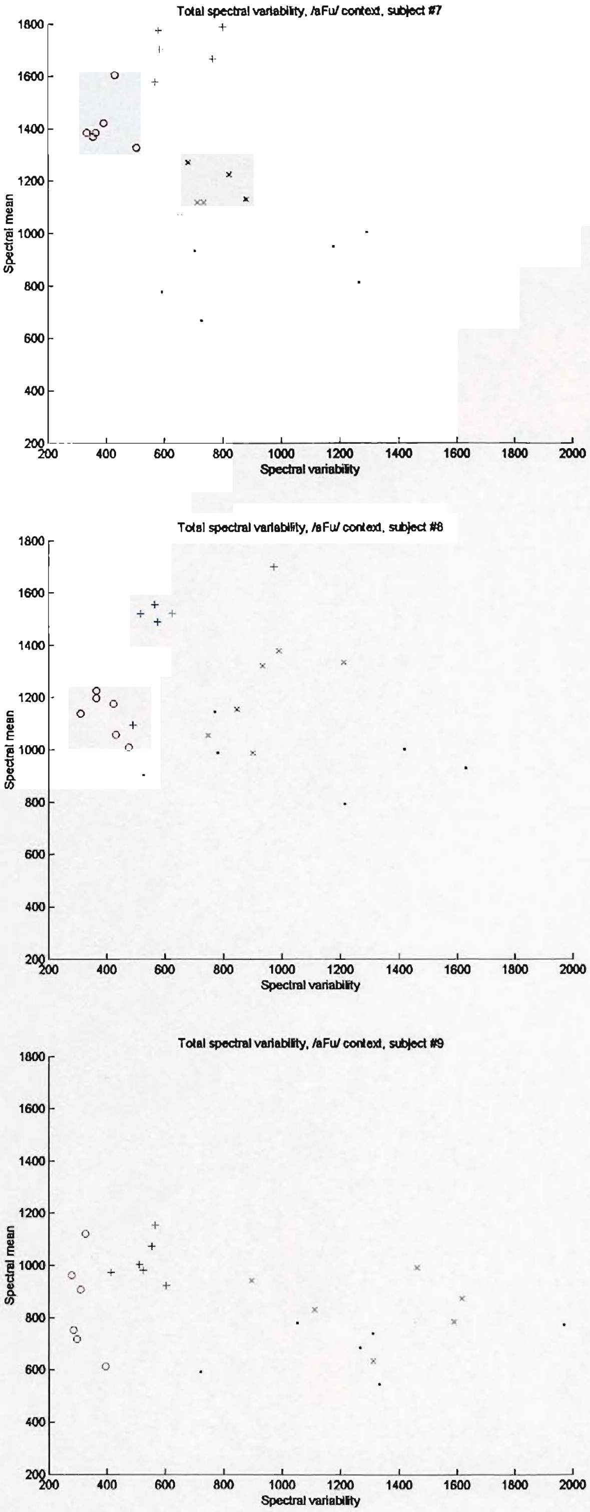


FIGURE F.71: Total spectral variability of all fricatives in /əFu/ context, subjects F-07 to F-09.

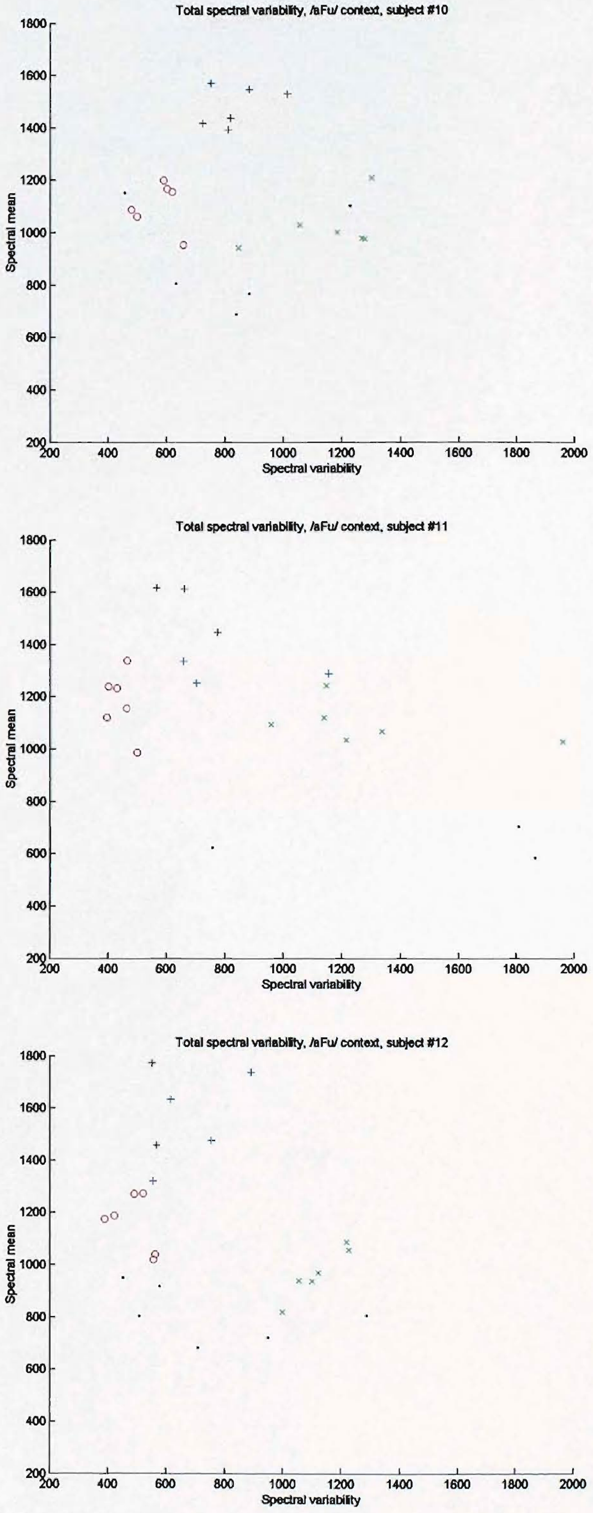


FIGURE F.72: Total spectral variability of all fricatives in /əFu/ context, subjects F-10 to F-12.

## F.5 Peak variability by subject and example vowel context

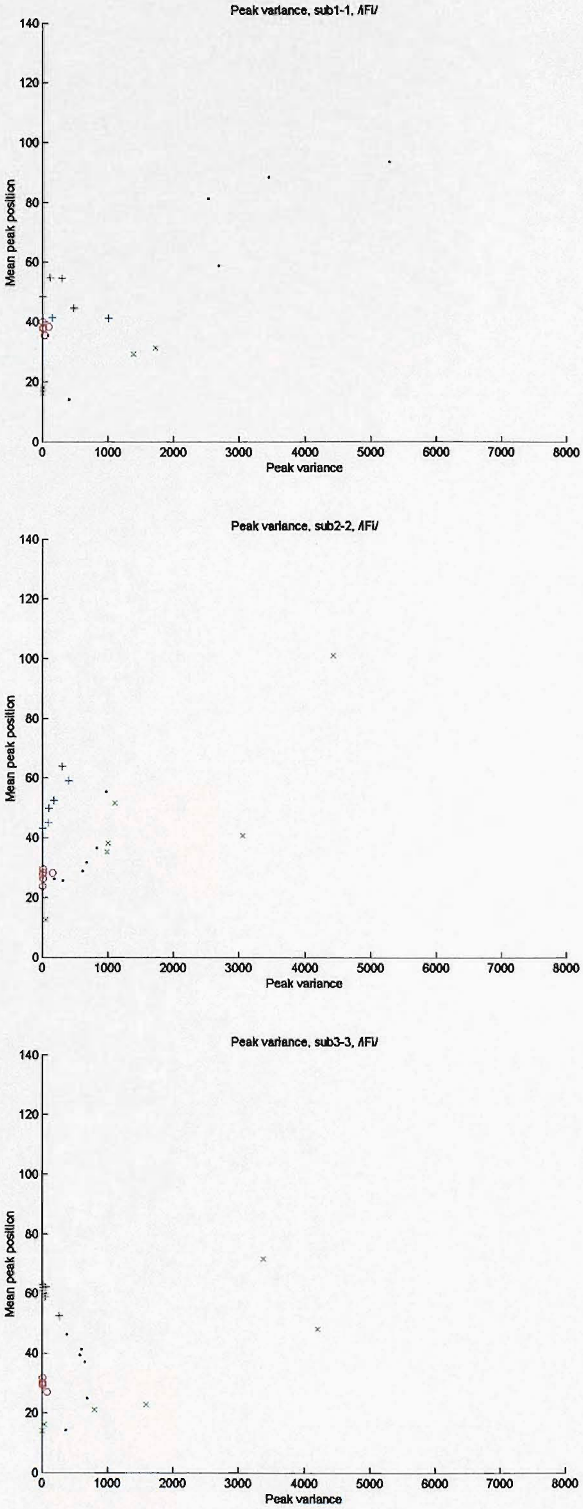
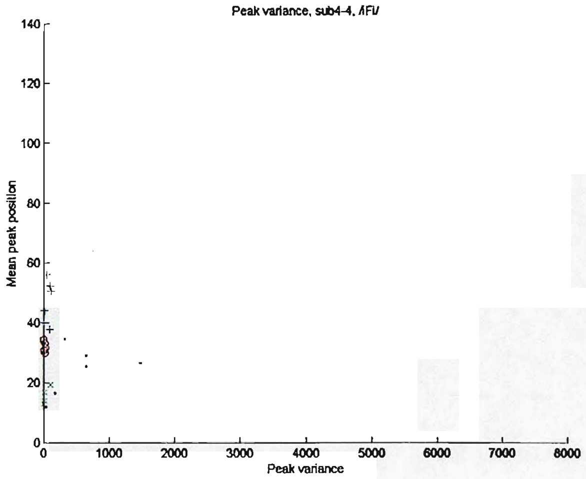
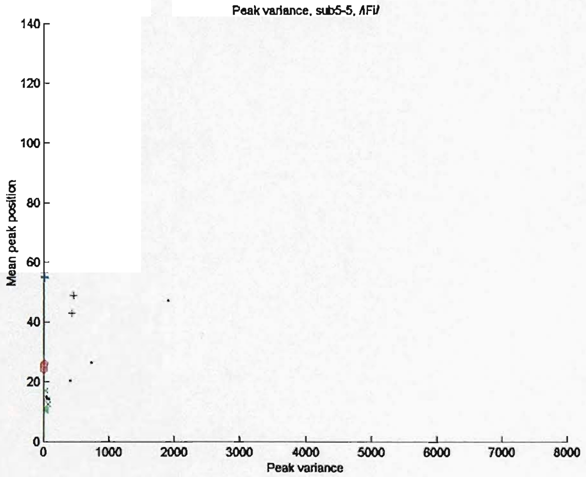


FIGURE F.73: Peak variability of all fricatives in /iFi/ context, subjects M-01 to M-03.





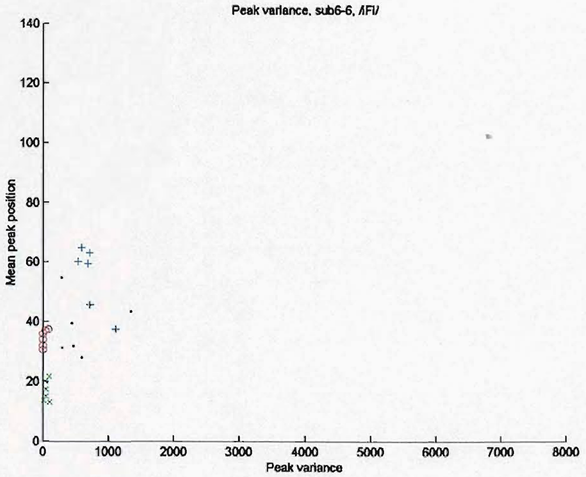


FIGURE F.74: Peak variability of all fricatives in /iFi/ context, subjects M-04 to M-06.

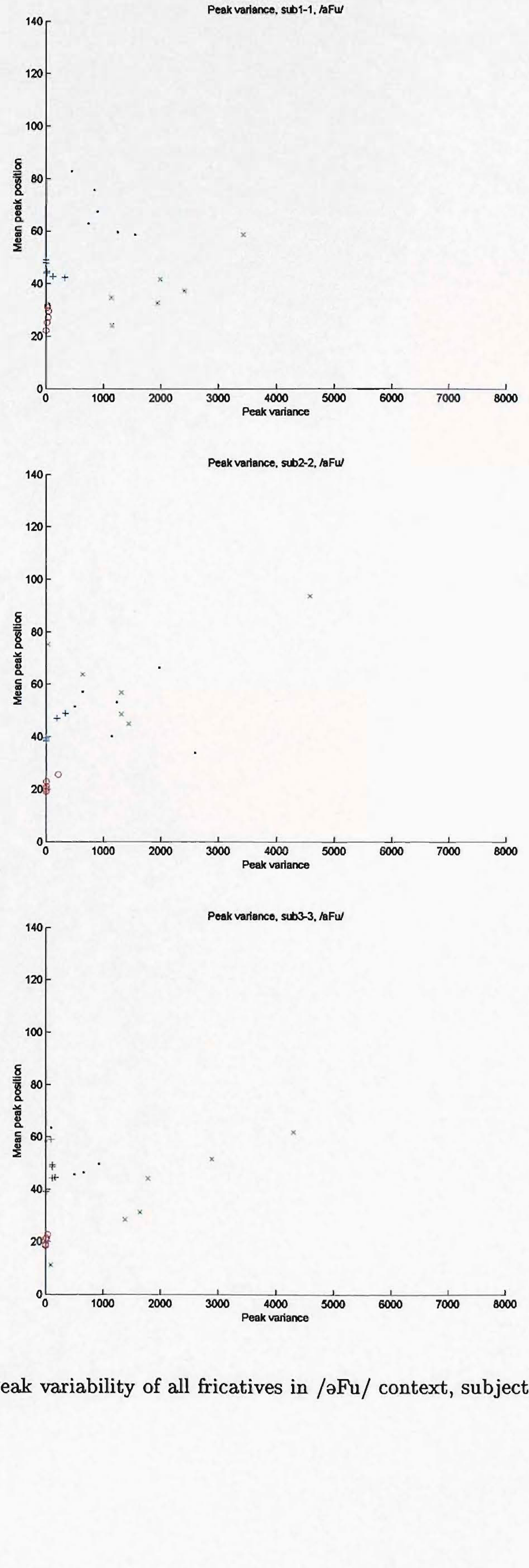
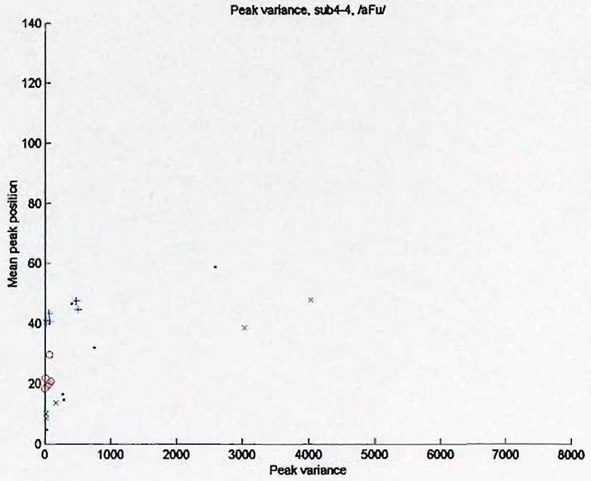
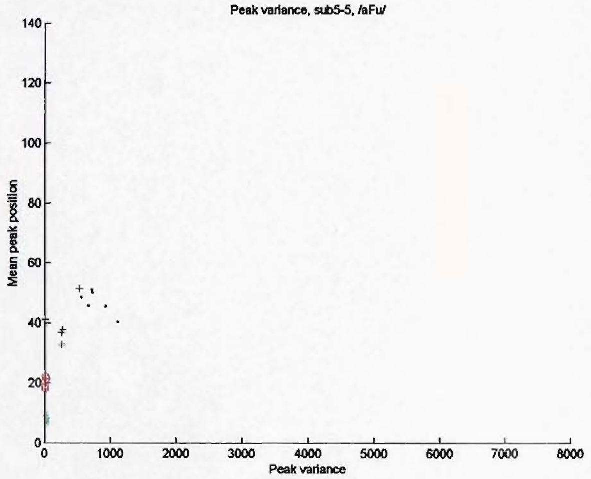


FIGURE F.75: Peak variability of all fricatives in /əFu/ context, subjects M-01 to M-03.





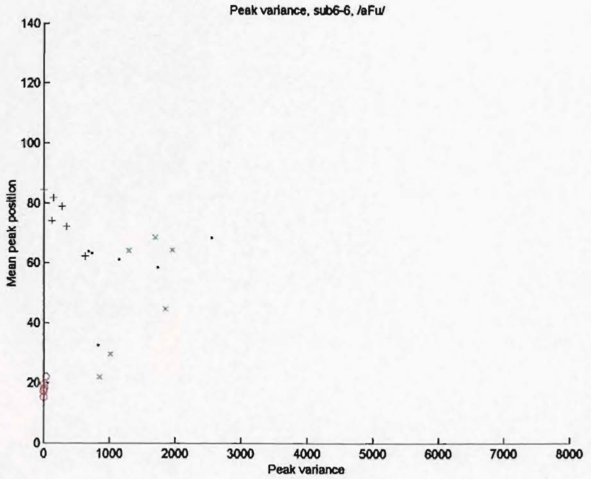


FIGURE F.76: Peak variability of all fricatives in /əFu/ context, subjects M-04 to M-06.

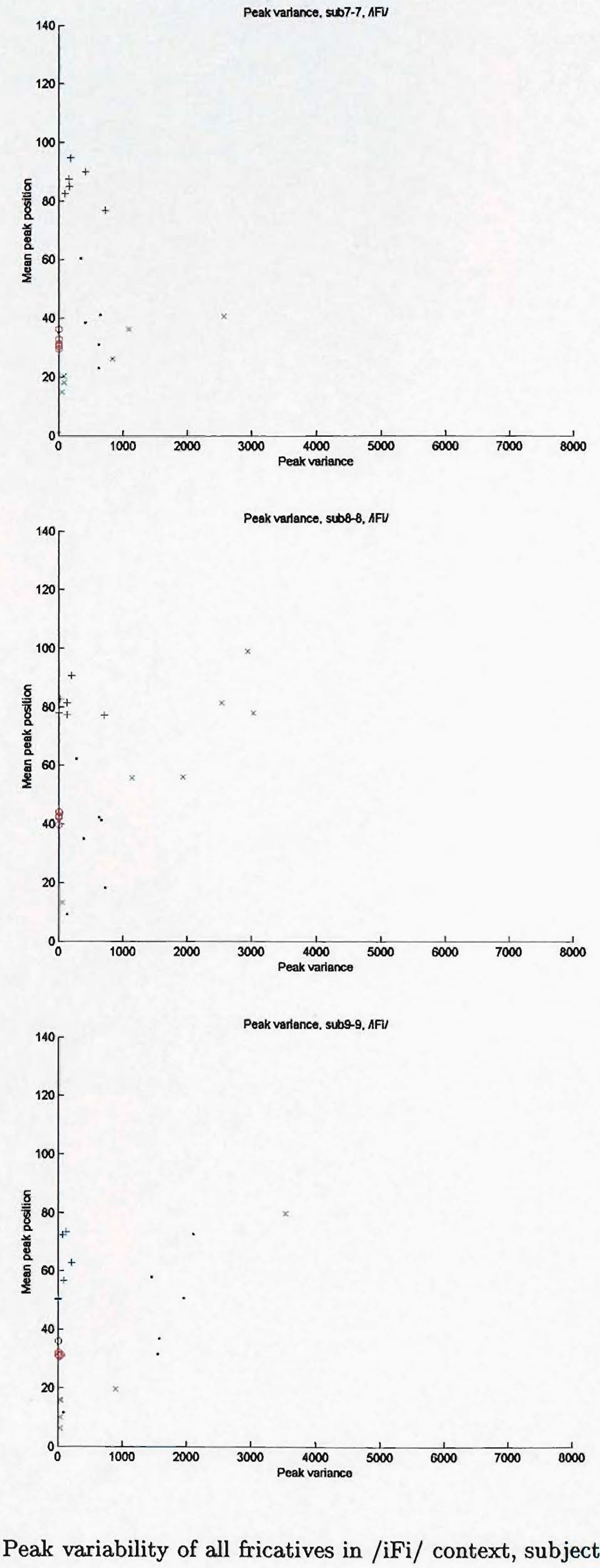


FIGURE F.77: Peak variability of all fricatives in /iFi/ context, subjects F-07 to F-09.

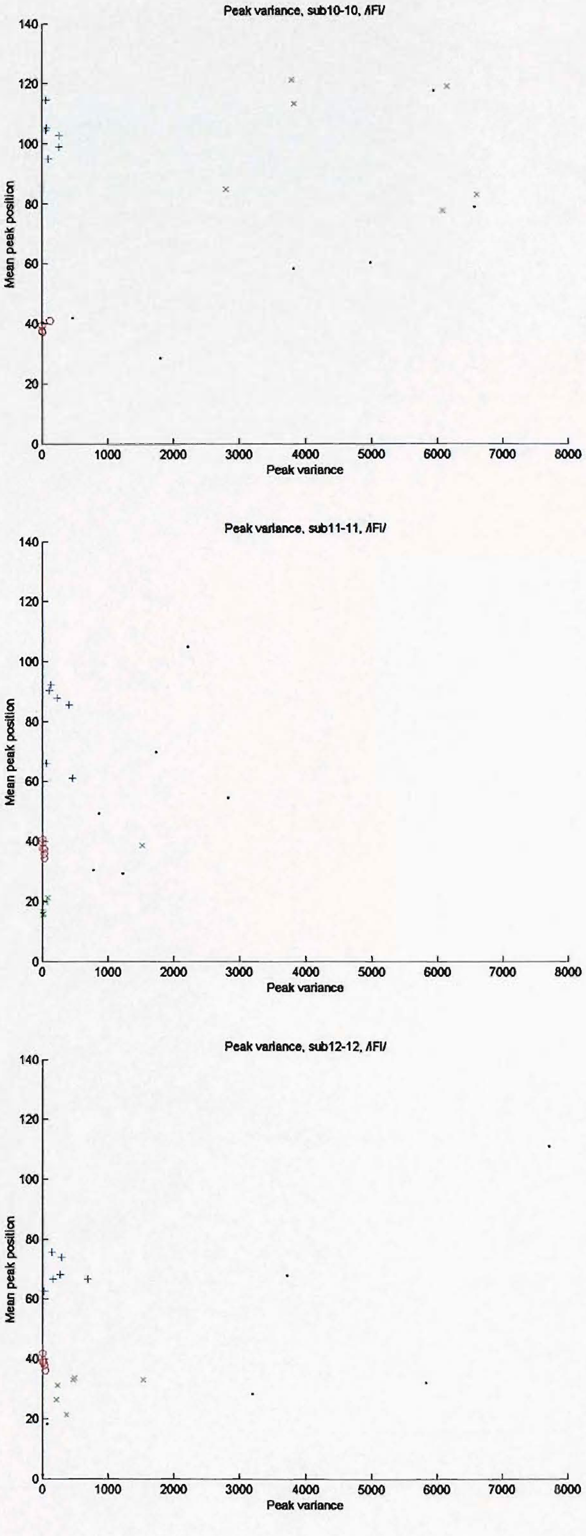


FIGURE F.78: Peak variability of all fricatives in /iFi/ context, subjects F-10 to F-12.

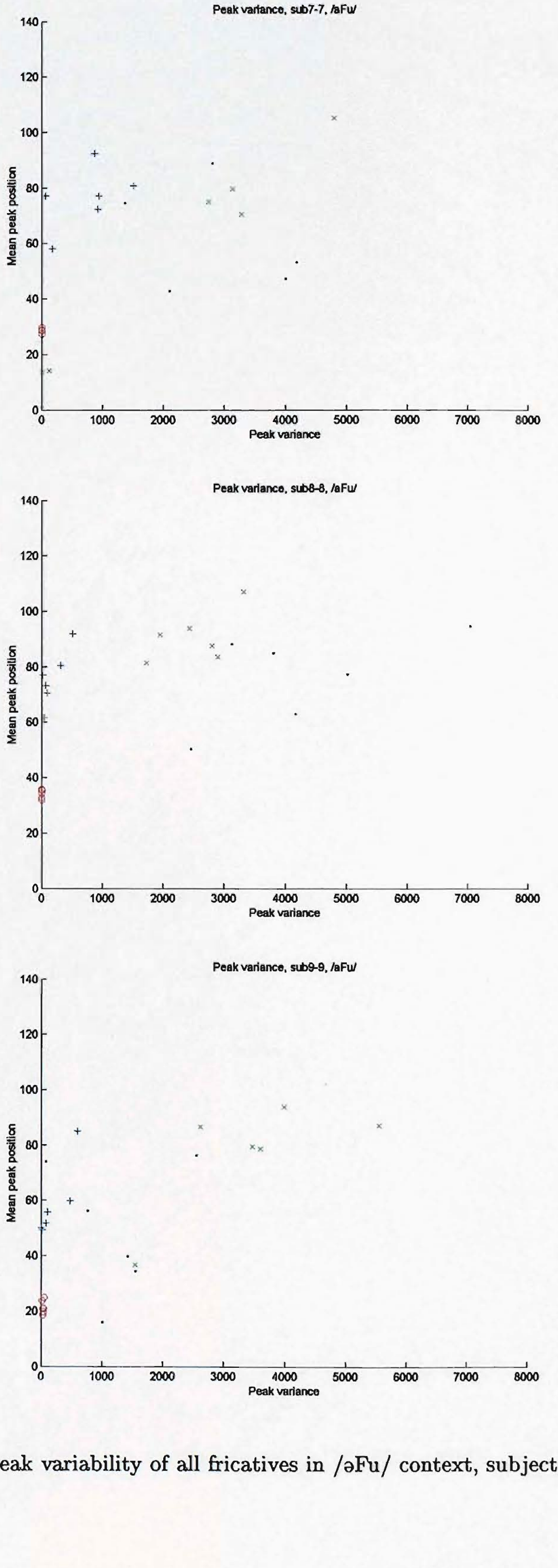


FIGURE F.79: Peak variability of all fricatives in /əFu/ context, subjects F-07 to F-09.

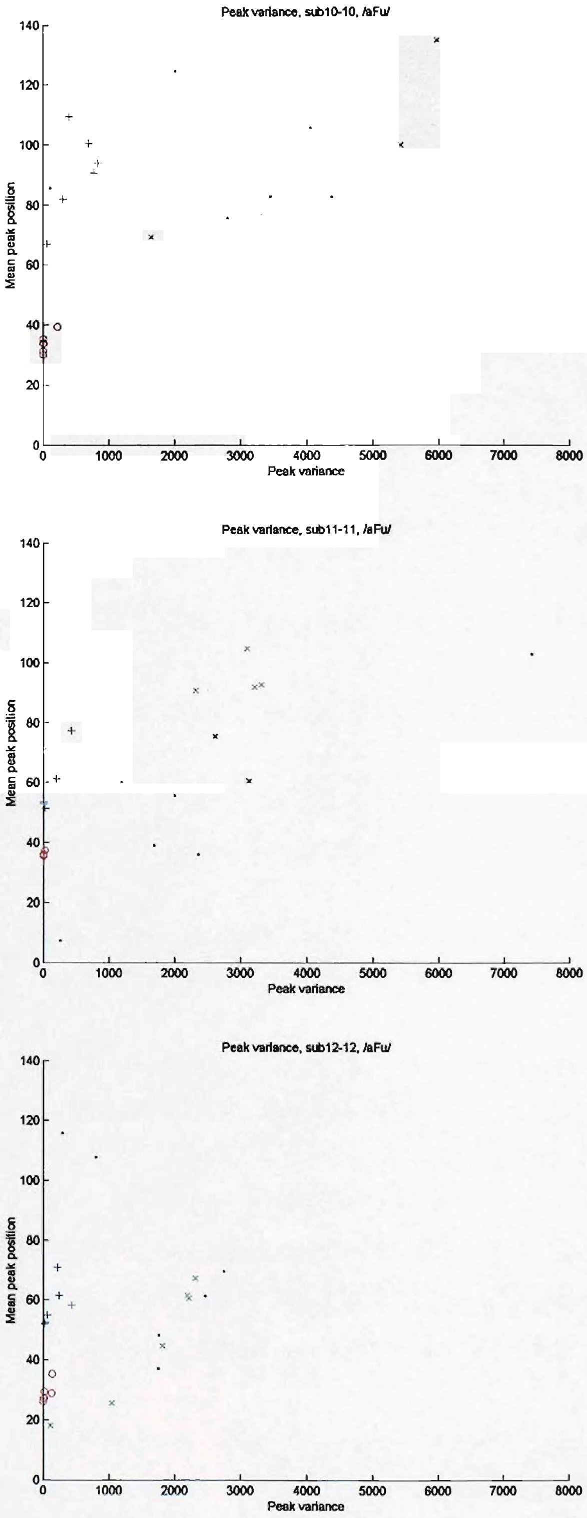


FIGURE F.80: Peak variability of all fricatives in /əFu/ context, subjects F-10 to F-12.

# Bibliography

Abbs, M. S. and F. D. Minifie (1969). Effect of acoustic cues in fricatives on perceptual confusions in preschool children. *J. Acoust. Soc. Am. 46*(6), 1535–1542.

Ali, A. M. A., J. Van der Spiegel, and P. Mueller (2001). Acoustic-phonetic features for the automatic classification of fricatives. *J. Acoust. Soc. Am. 109*(5), 2217–2236.

Baum, S. R. and S. E. Blumstein (1987). Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *J. Acoust. Soc. Am. 82*(3), 1073–1077.

Baum, S. R. and J. C. McNutt (1990). An acoustic analysis of frontal misarticulation of /s/ in children. *J. Phon. 18*, 51–63.

Bendat, J. S. and A. G. Piersol (1986). *Random Data: Analysis and measurement procedures* (2nd ed.). Wiley-Interscience.

Blacklock, O. S. and C. H. Shadle (2003). Spectral moments and alternative methods of characterizing fricatives. In *Proc. 145th Meeting of the Acoustical Society of America*, Nashville, TN, pp. 2199.

Blumstein, S. E. and K. N. Stevens (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am. 66*(4), 1001–1017.

Cheesman, M. F. and K. G. Greenwood (1995). Selective adaptation by context-conditioned fricatives. *J. Acoust. Soc. Am. 97*(1), 531–538.

Crystal, T. H. and A. S. House (1988). A note on the durations of fricatives in American English. *J. Acoust. Soc. Am. 84*(5), 1932–1935.

Fairbanks, G. (1960). *Voice and articulation drillbook.* New York: Harper & Row.

Fant, G. (1970). *Acoustic theory of speech production* (2nd ed.). The Hague, Paris: Mouton.

Flanagan, J. L. (1972). *Speech analysis synthesis and perception* (2nd ed.). Würzburg: Springer-Verlag.

Fletcher, S. G. and D. G. Newman (1991). [s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width. *J. Acoust. Soc. Am. 89*(2), 850–858.

Formby, C. and D. G. Childers (1996). Labelling and discrimination of a synthetic fricative continuum in noise: a study of absolute duration and relative onset time cues. *J. Speech Hearing Res. 39*, 4–18.

Forrest, K., G. Weismer, P. Milenkovic, and R. N. Dougall (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am. 84*(1), 115–123.

Goldstein, M. E. (1976). *Aeroacoustics.* New York: McGraw-Hill.

Gurlekian, J. A. (1981). Recognition of the spanish fricatives /s/ and /f/. *J. Acoust. Soc. Am. 70*(6), 1624–1627.

Harris, K. S. (1954). Cues for the identification of the fricatives of American English. *J. Acoust. Soc. Am. 26*, 952.

Hata, K., H. Moran, and S. Pearson (1994). Distinguishing the voiceless fricatives F and TH in English: a study of relevant acoustic properties. In *Proc. International Conference on Spoken Language Processing 94*, Yokohama, Japan, pp. 327–330.

Hawkins, S. (1999a). Looking for invariant correlates of linguistic units: two classical theories of speech perception. In J. M. Pickett (Ed.), *The acoustics of speech communication: fundamentals, speech perception theory, and technology*, pp. 198–231. Needham Heights, MA.: Allyn and Bacon.

Hawkins, S. (1999b). Reevaluating assumptions about speech perception: interactive and integrative theories. In J. M. Pickett (Ed.), *The acoustics of speech communication: fundamentals, speech perception theory, and technology*, pp. 232–288. Needham Heights, MA.: Allyn and Bacon.

Hedrick, M. S. and R. N. Ohde (1993). Effect of relative amplitude of frication on perception of place of articulation. *J. Acoust. Soc. Am. 94*(4), 2005–2026.

Heinz, J. M. and K. N. Stevens (1961). On the properties of voiceless fricative consonants. *J. Acoust. Soc. Am. 33*(5), 589–596.

Hughes, G. W. and M. Halle (1956). Spectral properties of fricative consonants. *J. Acoust. Soc. Am. 28*(2), 303–310.

Jassem, W. (1979). Classification of fricative spectra using statistical discriminant functions. In B. Lindblom and S. Öhman (Eds.), *Frontiers of Speech Communication Research*, pp. 77–91. Academic Press.

Jesus, L. M. T. and C. H. Shadle (2002). A parametric study of the spectral characteristics of european portugese fricatives. *J. Phonetics 30*(3), 437–464.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Lang. and Speech. 34*(3), 265–279.

Jongman, A. (1989). Duration of fricative noise required for identification of English fricatives. *J. Acoust. Soc. Am. 85*(4), 1718–1725.

Jongman, A. and J. A. Sereno (1995). Acoustic properties of non-sibilant fricatives. In *Proc. International Congress of Phonetic Sciences 95*, Stockholm, Sweden, pp. 432–435.

Jongman, A., R. Wayland, and S. Wong (2000). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am. 108*(3), 1252–1263.

Kenney, J. K. and E. S. Keeping (1964). *Mathematics of Statistics* (3rd ed.), Volume 2. D. Van Nostrand Company, Inc.

Kent, R. D. and F. D. Minifie (1977). Coarticulation in recent speech production models. *J. of Phonetics 5*, 115–133.

Lacerda, F. P. (1982). Acoustic perceptual study of the Portugese voiceless fricatives. *J. Phonetics 10*, 11–22.

Landahl, M. T. (1975). Wave mechanics of boundary layer turbulence and noise. *J. Acoust. Soc. Am. 57*(4), 824–831.

Lane, H. and J. W. Webster (1991). Speech deterioation in postlingually deafened adults. *J. Acoust. Soc. Am. 89*(2), 859–866.

Lane, H., J. Wozniak, and J. Perkell (1994). Changes in voice-onset time in speakers with cochlear implants. *J. Acoust. Soc. Am. 96*(1), 56–64.

Lindblom, B. (1983). Economy of speech gestures. In P. N. MacNeilage (Ed.), *The Production of Speech*, pp. 217–245. Springer Verlag.

Lippmann, R. P. (1996). Accurate consonant perception without mid-frequency speech energy. *IEEE Trans. Speech Audio Proc. 4*(1), 66–69.

Löfqvist, A. (1999). Interarticulator phasing, locus equations, and degree of coarticulation. *J. Acoust. Soc. Am. 106*(4), 2022–2030.

Matthies, M. L., M. Svirsky, J. Perkell, and H. Lane (1996). Acoustic and articulatory measures of sibilant production with and without auditory feedback from a cochlear implant. *J. Speech Hear. Res. 39*, 936–946.

Matthies, M. L., M. A. Svirsky, H. L. Lane, and J. S. Perkell (1994). A preliminary study of the effects of cochlear implants on the production of sibilants. *J. Acoust. Soc. Am. 96*(3), 1367–1373.

McGowan, R. S. and S. Nittrouer (1988). Differences in fricative production between children and adults: evidence from an acoustic analysis of /ʃ/ and /s/. *J. Acoust. Soc. Am. 83*(1), 229–236.

Meyer-Eppler, W. (1953). Zum erzeugungsmechanismus der geräuschlaute. *Z. für Phonetik 7*(3), 196–212. Translated by Hecker, M.

Munson, B. (2001). A method for studying variability in fricatives using dynamic measures of spectral mean. *J. Acoust. Soc. Am. 110*(2), 1203–1206.

Newman, R. S., S. A. Clouse, and J. L. Burnham (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am. 109*(3), 1181–1196.

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *J. Acoust. Soc. Am. 97*(1), 520–530.

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *J. Acoust. Soc. Am. 112*(2), 711–719.

Nittrouer, S. and M. Studdert-Kennedy (1987). The role of coarticulation effects in the perception of fricatives by children and adults. *J. Speech Hearing Res. 30*, 319–329.

Nittrouer, S., M. Studdert-Kennedy, and R. S. McGowan (1989). The emergence of phonetic segments: evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *J. Speech Hearing Res. 32*, 120–132.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine.* Addison-Wesley.

Parker, A. (1999). *PETAL Phonological Evaluation & Transcription of Audio-Visual Language.* Chesterfield, UK: Winslow.

Percival, D. B. and A. T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques.* Cambridge, UK: Cambridge University Press.

Priestley, M. B. (1999). *Spectral Analysis and Time Series*, Volume 1 & 2 of *Probability and Mathematical Statistics*. London, UK: Academic Press.

Schlichting, H. (1960). *Boundary Layer Theory* (4th ed.). Series in Mechanical Engineering. London: McGraw-Hill Book Company, Inc.

Scully, C. (1990). Articulatory synthesis. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Dordrecht, Netherlands: Kluwer Academic Publishers.

Scully, C. and E. Allwood (1985). Production and perception of an articulatory continuum for fricatives of English. *Speech Communication 4*, 237–245.

Scully, C., E. Castelli, E. Brearley, and M. Shirt (1992). Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives. *J. Phon. 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*. PhD thesis, Massachusetts Institute of Technology, Research Laboratory of Technology, Cambridge, MA 02139. RLE TR-506.

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. Dordrecht, Netherlands: Kluwer Academic Publishers.

Shadle, C. H. and S. J. Mair (1996). Quantifying spectral characteristics of fricatives. In *Proc. International Conference on Spoken Language Processing 96*, Philadelphia, PA, pp. 1521–1524.

Shadle, C. H., A. Moulinier, C. U. Dobelke, and C. Scully (1992). Ensemble averaging applied to the analysis of fricative consonants. In *Proc. International Conference on Spoken Language Processing 92*, pp. 53–56.

Shadle, C. H. and C. Scully (1995). An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. *J. Phon. 23*, 53–66.

Slepian, D. (1978). Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: the discrete case. *Bell System Tech. J. 43*, 3009–3057.

Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *J. Acoust. Soc. Am. 70*(4), 976–984.

Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am. 50*(4), 1180–1192.

Stevens, K. N. (1989). On the quantal nature of speech. *J. Phon. 17*, 3–45.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N., S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *J. Acoust. Soc. Am. 91*(5), 2979-3000.

Strevens, P. (1960). Spectra of fricative noise in human speech. *Language and Speech 3*, 32–49.

Sussman, H. M., H. A. McCaffrey, and S. A. Matthews (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am. 90*(3), 1309–3125.

Sussman, H. M. and J. Shore (1996). Locus equations as phonemic descriptors of consonantal place of articulation. *Perception and Psychophysics 58*(6), 936–946.

Syrdal, A. K. and H. S. Gopal (1986). A perceptual model of vowel recognition based on auditory representation of American English vowels. *J. Acoust. Soc. Am 79*, 1086–1100.

Thomson, D. (2000). Multitaper analysis of nonstationary and nonlinear time series data. In W. Fitzgerald, R. Smith, A. Walden, and P. Young (Eds.), *Nonlinear and Nonstationary Signal Processing* (1st ed.)., Chapter 11, pp. 317–394. Cambridge, UK: Cambridge University Press.

Van Dyke, M. (1982). *An Album of Fluid Motion.* Stanford, CA: Parabolic Press.

Vandali, A. E. (2001). Emphasis of short-duration acoustic speech cues for cochlear implant users. *J. Acoust. Soc. Am. 109*(5), 2049–2061.

Weinstein, C. J., S. G. McCandless, L. E. Mondschein, and V. W. Zue (1975). A system for acoustic-phonetic analysis of continuous speech. *Trans. IEEE-ASSP 23*(1), 54–67.

Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. on Audio and Electroacoustics AU-15*(2), 70–73.

Whalen, D. H. (1991). Perception of the English /s/-/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *J. Acoust. Soc. Am. 90*(4), 1776–1785.

Wilson, B. S. (1993). Signal processing. In R. S. Tyler (Ed.), *Cochlear Implants: Audiological Foundations*, Chapter 2, pp. 35–85. San Diego, CA: Singular Publishing Group, Inc.

Wrench, A. A. (1995). Analysis of fricatives using multiple centres of gravity. In *Proc. International Congress of Phonetic Sciences 95*, Volume 4, Stockholm, Sweden, pp. 460–463.

Yeni-Komshian, G. H. (1981). Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. *J. Acoust. Soc. Am. 70*(4), 966–975.

Zeng, F. and C. W. Turner (1990). Recognition of voiceless fricatives by normal and hearing-impaired subjects. *J. Speech Hearing Res. 33*, 440–449.