Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification

by

Mark Allan Hasegawa-Johnson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1996

Author ....
Department of Electrical Engineering and Computer Science
August 15, 1996

Certified by ....
Kenneth N. Stevens
Clarence J. LeBel Professor of Electrical Engineering
Thesis Supervisor

Accepted by ....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification

by
Mark Allan Hasegawa-Johnson

Submitted to the Department of Electrical Engineering and Computer Science
on August 15, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

This thesis demonstrates that acoustic variability, acoustic measurement error, and phoneme classification error can be interpreted as predictable entailments of articulatory variability. Speech production theory is tapped to explain sources of variability in the acoustic signal, including random variation in a turbulent spectrum, increased losses at the glottis, and coloration of the spectrum by subglottal and back cavity resonances. Measurements of the burst front cavity resonance, and of formant frequencies, which are defined as the eigenfrequencies of the vocal tract, are developed using both knowledge-based and HMM design methods, and are evaluated using the tools of acoustic phonetics and of statistical speech classification.

The error or uncertainty of both rule-based and HMM algorithms is evaluated by comparison to the measurements of human judges on a test set. Measurement error of the rule-based algorithm is evaluated using aggregate statistical models, including explicit models of outliers and heteroskedasticity, and a non-parametric model of the effect on error of phonetic context. Measurement uncertainty of the HMM formant tracker is calculated by the HMM itself during the measurement process. The uncertainty models generated by the HMM formant tracker are compared to formants transcribed by human judges, and shown to provide imperfect but generally acceptable predictions of the measurement error.

Acoustic variability and acoustic measurement error are evaluated using the tools of phonetic classification. Context-independent linear discriminant classification of stop place, using tokens from the TIMIT multi-speaker database, is shown to be 89% correct using manual formant and burst spectral measurements, but only 76% correct using automatic measurements. It is demonstrated that the difference between the classification of manual and automatic measurements can be accurately predicted using a heteroskedastic error model. Context-dependent classification experiments using both rule-based and HMM formant measurements result in 83-84% correct classification over the TIMIT TEST database. The pattern of classification errors as a function of phonetic context is shown to be similar to the pattern of errors of human listeners, indicating that the types of acoustic variability which confuse the classifier may be similar to the types of variability which confuse human listeners.

Thesis Supervisor: Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering

Acknowledgments

A locksmith fixing my car door once billed me ten dollars for a curved piece or metal smaller than my fingernail. "How can something so small be so expensive?" I asked. "Of course it's small," he replied, "but nothing else will do the job."

Many people have contributed to this thesis. Some contributions are small, and some are large, but all are irreplaceable.

Several people deserve credit for introducing me to important papers which I wouldn't have found on my own. Victor Zue introduced me to the works of Lori Lamel and of Gary Kopec, both of which have had tremendous influence on this thesis. Jonathan Allen loaned me his copy of Tukey's *Exploratory Data Analysis*, which helped to shape the philosophy of this thesis, even though I never found occasion to cite it. John Ohala referred me to the paper by Winitz, Scheib, and Reeds, and suggested the use of sound change evidence as a proxy for perceptual confusions. Finally, the papers by Halle, Hughes, and Radley, and by Nossair and Zahorian, were suggested by a helpful observer at the 1995 ASA meeting in St. Louis, who escaped before I got a chance to look at his name tag.

Drew Halberstadt deserves credit for running Raymond Chun's classification scripts for me, and for extracting the data shown in table 1.1. Jim Glass deserves credit for suggesting the idea; Jim's suggestions are always good ones.

MIT gave me lots of advice about how to finish my doctoral thesis, but most of the best advice (things like "don't delete things — just copy them to a clipboard file") came from those who made it through ahead of me. Particularly useful suggestions have come from Corine Bickley, Abeer Alwan, Lorin Wilde, Helen Hanson, Hwa-Ping Chang, and Sharlene Liu, among others. Walter Sun deserves credit for teaching me much of what I know about matlab, and for enjoyable statistical analysis jam sessions.

Tim Hazen, Krishna Govindarajan, and Wil Howitt deserve credit for dragging me out for beer occasionally (although drinking the night before my defense was my own idea!). Marilyn Chen and Jeff Kuo are party animals in their own way, and helped to make MIT a nicer place. Stefanie Shattuck-Hufnagel, Arlene Wint, and others in the Speech Communication Group helped contribute to a supportive environment in which it was a joy to be a student.

My readers, Jonathan Allen and Victor Zue, gave me many good ideas, of which I had time to implement only a small fraction. Both deserve credit, most of all, for believing in the importance of my thesis even when I myself did not, for refusing to let me bluff when I didn't know an answer, and for refusing to let me turn in a mediocre document when they believed it could be better.

All of the people I have mentioned above contributed important parts to this puzzle, but most of the credit for the hundred-odd pages you are about to read belongs to four outstanding individuals.

To all those who claim that professors sponsor students because they have to, or in order to advance their own carcers, I offer Ken Stevens as a counter-example. Professor Stevens cares deeply about his students, and I have often seen him go far out of his way to provide for their mental, emotional, and physical well-being. In a profession where most love either teaching or research, Ken loves both. In a profession in which the pressure to specialize is tremendous, Ken has made a name for himself by bringing the tools of electrical engineering to bear on the problems of linguistics. He is an unsurpassed mentor, and an admirable role model.

If it weren't for the values and skills taught by my father and mother, I would never

have made it to MIT in the first place. My parents taught me the value of education — not monetary value, but personal and spiritual value. They taught me the joy of reading and learning. They taught me how to love, and how to be honest with myself, and how important each is for the other. Most of all, they taught me that "just because" is never an acceptable answer. Everything has an explanation; some explanations just take a little more time to understand.

Finally, I would like to thank my wife, Yu, and our cats Nyan Sama, Funny Face, and Baby Torachan, for making my life happy. Yu keeps me up at night for hours arguing about the nature of intelligence, and then videotapes me playing with the kitten. She edits my research proposals, and suggests important corrections, even when half of the words are in a jargon she doesn't understand. She not only believes in me more than I do myself, but she brags about me to her friends and family until they get sick of hearing about it. Without Yu's encouragement and enthusiasm, I would never have had the confidence to propose, execute, or report the thesis you hold in your hands.

This work has been supported in part by grants from the National Institute on Deafness and other Communication Disorders (NIDCD) and the National Science Foundation (NSF).

Contents

1 Introduction: The Acoustic Correlates of Place **16**

1.1 Measurements Used in this Thesis 17

1.2 Previous Studies: Automatic Formant Tracking 21

1.2.1 An HMM formant model 21

1.2.2 Finding the roots of the LPC polynomial 22

1.3 Previous Studies: Acoustic Cues for Place Classification 22

1.3.1 Onset spectral measurements for classification 23

1.3.2 Formant frequency information 23

1.3.3 Dynamic spectral information 24

1.3.4 Classification using TIMIT 24

1.4 Discussion: Acoustic Correlates of Place 25

1.5 Discussion: Classification as a Tool for the Analysis of Variability 26

1.6 Thesis Outline 27

2 Speech Production Theory **29**

2.1 Speech Sources 30

2.1.1 Transient source 30

2.1.2 Turbulent sources 31

2.1.3 Glottal vibration 34

2.1.4 Measurement issues: time-averaged power spectrum 34

2.2 Speech Filters 37

2.2.1 All-pole models 37

2.2.2 Factors which influence formant bandwidth 38

2.2.3 Pole-zero pairs 40

2.2.4 Nasalization 43

2.2.5 Measurement issues: LPC 44

2.2.6 Measurement issues: frication spectrum 46

2.3 Evolution of Vocal Tract Resonances at Release of a Consonant 48

2.3.1 First formant trajectory 48

2.3.2 Labial releases 49

2.3.3 Velar releases 50

2.3.4 Alveolar, retroflex, and lateral releases 51

2.4 Summary and Discussion 52

2.4.1 Acoustic correlates of place 52

2.4.2 Production models of variability 53

3 Rule-Based Measurement Algorithms 55
3.1 Signal Representation . . . 55
3.1.1 LPC-based formant frequencies . . . 56
3.1.2 Time-averaged power spectrum of the burst . . . 56
3.2 Knowledge Representation and Algorithm Design . . . 56
3.2.1 Burst front cavity resonance . . . 56
3.2.2 Other burst spectral measurements . . . 59
3.2.3 Formant frequencies above F1 . . . 60
3.2.4 First formant measure . . . 62
3.3 Imitating Human Performance on a Training Set . . . 64
3.3.1 Training data . . . 64
3.3.2 Formant measurement algorithms . . . 64
3.3.3 Burst spectral measurement algorithms . . . 65
3.4 Statistical Models of Measurement Error . . . 67
3.4.1 Reference measurements . . . 67
3.4.2 Confidence limits for a Gaussian error model . . . 68
3.4.3 Explicit modeling of outliers . . . 70
3.4.4 The effect of context on outlier probability . . . 72
3.4.5 Correlations between measurements and measurement errors . . . 74
3.5 Discussion . . . 76
3.5.1 Algorithm training . . . 76
3.5.2 Aggregate error models . . . 77

4 Stochastic Formant Modeling 82
4.1 Phoneme-Dependent HMM Formant Tracking . . . 82
4.1.1 A phoneme-dependent transition model . . . 82
4.1.2 A DFT-based output model . . . 84
4.2 Estimating Parameters from a Reference Transcription . . . 86
4.2.1 Means as a function of time . . . 86
4.2.2 Variance and covariance . . . 87
4.3 Real-Time Estimation of Measurement Uncertainty . . . 88
4.3.1 Finding a posteriori formant distributions . . . 89
4.3.2 Examples . . . 89
4.4 Evaluating the Error Models . . . 92
4.4.1 Cumulative-probability representation of reference measurements . . 92
4.4.2 Evaluation using a binomial distribution . . . 93
4.5 Conclusions . . . 95

5 Classification of Place 96
5.1 Effect of Measurement Error on the Linear Discriminant . . . 96
5.1.1 Measuring the effect of error . . . 97
5.1.2 The form of the discriminant error term . . . 99
5.1.3 Estimating confusion matrices from the discriminant error . . . 99
5.2 Classification using Rule-Based Measurements . . . 102
5.2.1 Optimizing free parameters . . . 103
5.2.2 Linear discriminant classification . . . 104
5.2.3 Comparison to the performance of human listeners . . . 106
5.2.4 Effect of right context . . . 107

5.3 Stochastic Formant Modeling 109
5.3.1 Modifications to the model 109
5.3.2 Example 111
5.3.3 Training from a large speech corpus 112
5.3.4 Classification of place 114
5.4 Summary 115

6 Conclusions **117**
6.1 Classification of Consonant Place 117
6.1.1 Context-independent classification 117
6.1.2 Context-dependent classification 118
6.2 Predicting the Effect of Measurement Error 119
6.2.1 Production models of the sources of error 119
6.2.2 Predicting classification on the basis of aggregate error models . . . 120
6.2.3 Real-time predictions of measurement uncertainty 121
6.3 Future Work: Stochastic Formant Models of Perception 121

A Speech Data **124**
A.1 KB Train: Training tokens for knowledge engineering 124
A.2 Error Modeling: Data used to develop error models 130
A.3 SFM Train: Training data for the HMM formant tracker 140

Bibliography **142**

List of Figures

1-1 F2 onset (first measurable F2) as a function of F2 target (40ms after consonant release), measured by human judges in 131 stop tokens from TIMIT. . 17

1-2 F3 onset (first measurable F3) as a function of F3 target (40ms after consonant release), measured by human judges in 131 stop tokens from the "test" subdirectory of TIMIT. 18

1-3 Composite F2 and F3 measurements, designed to differentiate lip and tongue places of articulation. 19

1-4 Amplitude of the highest peak in the F2-F6 range of the burst spectrum, plotted against a composite formant measurement designed to separate lip and tongue stops. 131 tokens, measured by human judges, from the "test" subdirectory of TIMIT. 20

1-5 Frequency of the front cavity resonance, as a function of F2 onset. 88 alveolar and velar stop releases, extracted from the "test" subdirectory of TIMIT, measured by human judges. 20

2-1 During the release of an unvoiced stop, four distinct acoustic sources are activated. The source contributions overlap in time; for example, ringing of the transient continues past the onset of frication. 30

2-2 Transmission line model of the vocal tract configuration immediately after release of a stop. The characteristic impedance of the vocal tract is Z_0, the constriction resistance is $R_c(t)$, the glottal impedance is set to infinity, and the radiation impedance is set to zero. The back cavity is charged to a nonzero pressure, which is equalized quickly as the admittance of the constriction grows. 31

2-3 Circuit models for the monopole and dipole turbulent acoustic sources. Radiation impedance and impedance of the back cavity are ignored; the mouth flow $U_m(t)$ is just the flow through a short circuit at one end of the transmission line. The monopole source $U_c(t)$ is in parallel with the constriction impedance; the dipole source $P_s(t)$ is connected in series, typically 1-3cm downstream from the constriction. 32

2-4 (a) Monopole turbulent source spectrum, based on the radiated spectra reported by Pastel (1987). (b) Flow source spectrum composed of the product of a dipole turbulent source, based on spectra reported by Shadle (1985), and a coupling factor $G_c(f)$ which assumes a vocal tract area of 5cm^2 and a distance of 2cm between constriction and source. Amplitudes reported by Pastel and Shadle have been adjusted to represent 300Hz bands, a flow of 420 cm^3/s, and a constriction area of 0.08 cm^2. 32

2-5 Schematized glottal flow waveform and spectrum. 34

2-6 Circuit used for calculating the contribution of constriction losses to bandwidth. Resistance R_c and inductance L_c are shown in parallel with a constriction source flow U_s; vocal tract is shown as a transmission line with characteristic impedance Z_0. 39

2-7 Circuit used for calculating the contribution of radiation losses to bandwidth. Radiation inductance L_r and resistance $R_r(f)$ are shown in series with the vocal tract transmission line. 40

2-8 Three-tube model of the vocal tract, and circuit models of the back and front cavities, during production of an /i/. L_g and R_g are the glottal resistance and inductance, L_c and R_c are the constriction resistance and inductance, and L_r and $R_r(f)$ are the radiation resistance and inductance. 41

2-9 Source-filter models of frication at the lips, and of frication at a tongue constriction. 43

2-10 Low-frequency lossless circuit model of the vocal tract after release of a stop, used for calculating the first formant frequency. The back cavity is treated as a lumped capacitor C_b, which resonates with the parallel inductances of the yielding wall, L_w, and stop constriction, L_c. 49

2-11 Simplified tube models of the vocal tract immediately after release of a labial stop, with different shapes of the back cavity. 49

2-12 Simplified tube models of the vocal tract immediately after release of a velar stop, with backed and fronted tongue positions. 50

2-13 Simplified tube models of the vocal tract immediately after release of alveolar and retroflex stops. 51

3-1 Schematized *a priori* distribution of possible front cavity resonance frequencies. F2(t+20) and F3(t+20) are the F2 and F3 measurements 20 milliseconds after release, which are assumed to be the first reliable formant measurements. With no knowledge of the stop or vowel identity, the most we can say is that the front cavity resonance must be between the vicinity of F2(t+20) and some fixed upper bound. 57

3-2 Schematized *a posteriori* probability distribution of the front cavity resonance frequency F_{f1}, given a particular burst spectrum. In the example shown, five convex spectral peaks have been identified in the frequency range of interest, with log amplitudes of 74, 60, 74, 67, and 40dB, respectively (measured relative to an arbitrary minimum amplitude). According to formulas specified in the text, the *a posteriori* probability of the front cavity resonance being located at any of these five peak frequencies is proportional to the log amplitude of the corresponding spectral peak. 59

3-3 Schematized *a priori* distribution of a given formant frequency, as a function of a later measurement of the same formant. With no information about the consonant or the vowel, we can only predict that the formant is continuous, hence a bell-shaped distribution. 61

3-4 Schematized *a posteriori* probability distributions for two formant frequency tracks. In the upper plot, the LPC root at time t is close to the formant at $t + \Delta t$, and is therefore judged to be a continuation of the formant. In the lower plot, the closest LPC root at time t is far from the formant at time $t + \Delta t$, and is therefore judged to be a measurement error. 62

3-5 Schematized *a priori* distribution of $F1$ at time t, given a measurement of $F1(t + \Delta t)$. Distribution is skewed, representing our expectation that F1 rises as a function of t. 63

3-6 Schematized *a posteriori* distributions of F1 at time t, given an LPC root $F_{r1}(t)$ which is (a) 75 Hertz below $F1(t+\Delta t)$, or (b) 75 Hertz above $F1(t+\Delta t)$. 63

3-7 Amplitudes of the highest peak in the F2/F3 and F4/F5/F6 range, velar (top) and non-retroflex alveolar (bottom) tokens only. 66

3-8 Q-Q plots describing the degree to which the measurement error fits a Gaussian model: if the errors fit a Gaussian model, the plot should be a straight line. Low-amplitude errors are approximately Gaussian, but there are many outliers. 69

3-9 The cumulative probability of finding a measurement error greater than any given frequency, with 95% confidence limits. Plots have been truncated at 500 Hertz, in order to make the threshold behavior near 150 Hertz more visible. 70

3-10 Mixture Gaussian models of the formant and front cavity resonance measurement errors (shown: log probability density). Formant mixtures were constrained to have the same mean; the front cavity resonance model was not so constrained. 72

3-11 Cumulative histograms, showing the probability of finding an error larger than the abscissa coordinate. The solid lines are the measured probabilities for low-error and high-error contexts; the dotted and dashed lines show 95% confidence limits for the low-error and high-error contexts, respectively. . . 73

3-12 Burst spectral amplitudes, as measured by human transcribers, and by the rule-based measurement algorithm. Measurements by human transcribers have been shifted by a constant in order to make the plots comparable. The rule-based measurement algorithm has a significant tendency to reduce high-amplitude measurements, and increase low-amplitude measurements. 76

4-1 Schematic representation of a phoneme-dependent hidden Markov model formant tracker. Time of the landmark is assumed given. The search space consists of a phoneme level, and a formant frequency level. Formant tracks are semi-continuous in frequency, and discrete in time, with a formant frequency space searched once every 10ms. Spacing of the circles in the schematized formant frequency search space, above, represents the height of the probability density curve as a function of frequency at each time. The output space consists of models of spectral amplitude and convexity as a function of formant frequency, phoneme identity, and time. Spectral amplitude and convexity are computed using appropriate transformations of the periodic time-averaged power spectrum, as described in the text. 83

4-2 Lifter used to estimate spectral convexity. The lifter is a seven-sample match to the spectrum of a 125Hz-bandwidth spectral resonance, normalized to zero mean and unit energy. The resulting lifter has a bandpass .epstrum, amplifying spectral features which have half-periods between roughly 67 and 167 Hertz. 85

4-3 Mean formant frequencies for three places of articulation, for each gender, trained on 36 tokens. Solid=labial, dashed=alveolar, dot-dash=velar. 87

4-4 Mean amplitude and convexity, averaged across all stop tokens, for three places of articulation. Solid=labial, dashed=alveolar, dot-dash=velar. . . . 88

4-5 Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 10ms and 50ms after release of the first /b/ in "Barb." *A posteriori* probability distributions of all three formants are shown on the same plot, because there is no significant overlap between the distributions. 90

4-6 Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 10ms and 50ms after release of the /d/ in "dark." *A posteriori* probability distributions of all three formants are shown on the same plot, because there is no significant overlap between the distributions. 91

4-7 Comparison of a known formant frequency to the uncertainty model generated by an HMM formant tracker. F2 is located at 1616 Hertz, according to the measurements of a human transcriber. In the uncertainty model generated by the HMM, the probability of an F2 measurement less than or equal to 1616 Hertz is 35.7%; thus we can say that the cumulative probability representation of the human transcription is $P = 0.357$. 93

4-8 Cumulative probability representations of the human formant measurements used to train the formant tracker (F1, F2, and F3). Deviation from a uniform distribution indicates that the formant tracker is not adequately representing variation in the training data. 94

4-9 Cumulative probability representations of formants measured by two human judges on an independent test set (F1, F2, and F3). Deviation from a uniform distribution indicates that the formant tracker is not adequately predicting variation in the new corpus. 95

5-1 LDA summary of measurements by human judges on 131 stop releases. The abscissa is a linear discriminant trained (on independent training data) to represent the labial/alveolar distinction; the ordinate is a discriminate trained to represent the labial/velar distinction. 97

5-2 LDA summary of automatic measurements of burst amplitude, burst peak count, and formant motion at 131 stop releases. The abscissa and ordinate are computed using the same linear discrimant coefficients applied to the data in figure 5-1. 98

5-3 Binary classifier. If both distributions are Gaussian as shown, the shaded area of the category 1 distribution is the probability that a category 1 token will be misclassified. 100

5-4 Example of three convex peak traces at the release of a /d/ by a female speaker. The left-hand plot shows the convex peak frequencies closest to expected values of F1, F2, and F3. The right-hand plot shows the amplitudes of these three peaks, and the amplitude AH of the largest peak higher in frequency than F3. 111

List of Tables

1.1 Stop place classification scores derived from the data of Chun (1996). The classifier is trained using the TIMIT TRAIN database (3696 sentences), and classification scores are reported for the TIMIT DEV database (400 sentences). 25

3.1 Signed differences between the two human judges on tokens which were transcribed by both. F1, F2, and F3 include six measurements per token, for more than 100 tokens. The burst spectral measures are the front cavity resonance frequency, low frequency and high frequency amplitudes, and relative diffuseness. 68

3.2 Mean and standard deviation of differences between automatic measurements and measurements transcribed by human judges, with bilateral 99% confidence limits. "Onset F1" includes measurements 0, 10, and 20ms after release, "Vowel F1" includes measurements 30, 40, and 50ms after release. . . 69

3.3 Separation of measurement errors into low-amplitude "normal errors" and high-amplitude "outliers." Measurements are F1, F2, and F3 (six measurements per consonant release) and the front cavity resonance (one measurement per stop release). 71

3.4 Definitions of seven frequency measurements, and of contexts in which the measurements are more likely to suffer peak-picking errors. High-error contexts were chosen on the basis of speech production theory, and of observation of the measured errors. 73

3.5 Normalized correlation between human transcriptions of the test data and measurement error of the automatic algorithm. 99% confidence limits are calculated using the Student's t distribution with 120 degrees of freedom. . 75

5.1 Percentage confusion matrices showing classification of vowel context stops in the Error Modeling database. The first confusion matrix shows classification using manual transcriptions; the second shows classification [illegible] automatic measurements. 98

5.2 Automatic measurement confusions predicted based on statistics of the manual measurements, and of the measurement error, using homoskedastic and heteroskedastic models. 101

5.3 Average magnitude difference, in each column, between the classification of automatic measurements, shown in table 5.1, and the predicted classification using homoskedastic and heteroskedastic error models, shown in table 5.2. . 102

5.4 Context-dependent measurement parameters were optimized independently for each combination of a speaker gender, a consonant manner class, and a right context. The 18 analyzed right contexts are shown here. Phoneme notation is ARPABET, as used in the TIMIT database. 103
5.5 The six band-edge frequencies shown were optimized, within the ranges shown, in order to minimize consonant place classification error in context-dependent LDA classifiers. 104
5.6 Confusion matrices using context-dependent formant and burst spectral measurements, categorized using a context-dependent LDA classifier. Stop tokens only, from the TEST subdirectory of TIMIT. All entries of 10% or less have been omitted. 105
5.7 Confusion matrices: linear discriminant classification of nasal consonants, using automatic formant measurements. TEST subdirectory of TIMIT. All entries of 15% or less have been omitted. 105
5.8 Classification of consonant place by human listeners, given full closure and release information (Lamel, 1988). Entries below 0.3% have been omitted. . 106
5.9 Classification of consonant place by human listeners, given release waveforms of varying length (Nossair and Zahorian, 1991). Entries less than 1% have been omitted. 106
5.10 Percentage classification of place as a function of right context. LB=low back vowels, HF=high front vowels and glides, HB=high back vowels and glides. Entries below 10% (LDA classifier) and 15% (human subjects) have been omitted. 108
5.11 Log-likelihood scores used to decide whether the token in figure 5-4 is a /d/ or a /g/. The table shows the difference in log-likelihood between /dah/ and /gah/ models for each formant frequency model and each formant amplitude, computed independently. 112
5.12 Confusion matrices for voiced stops, classified using the stochastic formant model. Entries of 10% or less have been omitted. 114
5.13 Stochastic formant model classification of voiced stops, as a function of right context. Entries of 10% or less have been omitted. 115

Chapter 1

Introduction: The Acoustic Correlates of Place

The sounds of English can be classified in terms of the quantized rotations, translations, and stiffnesses of about six articulators: the lips, tongue blade, tongue body, soft palate, pharynx, and vocal folds (Chomsky and Halle, 1968). Every phonetic distinction which enters the acoustic signal is placed there by the actions of these six or so articulators. It follows that a computer which is capable of tracking the salient changes in these six articulators over time should be capable of recognizing most of the linguistic content of an utterance.

Not all changes in the vocal tract shape, however, are linguistically salient. A listener (who can't see the speaker) has no information about tongue and lip positions except what he can get from the formant frequencies and amplitudes during a vowel or glide, or the front cavity resonances during a stop release or fricative. Since this is the only information about the tongue and lips available to the listener, it follows, again, that a computer capable of tracking the formants and front cavity resonances over time should be capable of recognizing most of the linguistic information carried by the tongue and lips, including consonant place and vowel quality.

The prospect of a compact, complete representation of the linguistic information in the signal prompted the development, in the 1970s and 1980s, of several speech recognizers based partly on formant analysis (see e.g. Klatt, 1977). All of these recognizers failed, in part, for the same reason: the formant tracking algorithms made mistakes, and the higher-level knowledge sources were unable to recover from the low-level mistakes. More recently, several phonetic studies have explored the possibility of using formant tracks to discriminate consonant place (Kewley-Port, 1982, Sussman et al., 1991) and vowel quality (Hillenbrand et al., 1995). These phonetic studies almost without exception rely on human judges to measure the formant frequencies, because automatic formant tracking algorithms are considered unreliable. Both the early recognition studies and the more recent phonetic studies assume that formant measurements must be perfect in order to be useful. Since the formant measurements produced by a tracking algorithm are never perfect, the algorithms have been judged to be useless.

This thesis proposes the use of imperfect automatic formant and front cavity resonance measurements as a tool in the analysis of phonetic variability. Phonetic classification (specifically, classification of consonant place) is used in this thesis as an experimental tool for analysis of the variability in automatic formant and burst spectral measurements, but

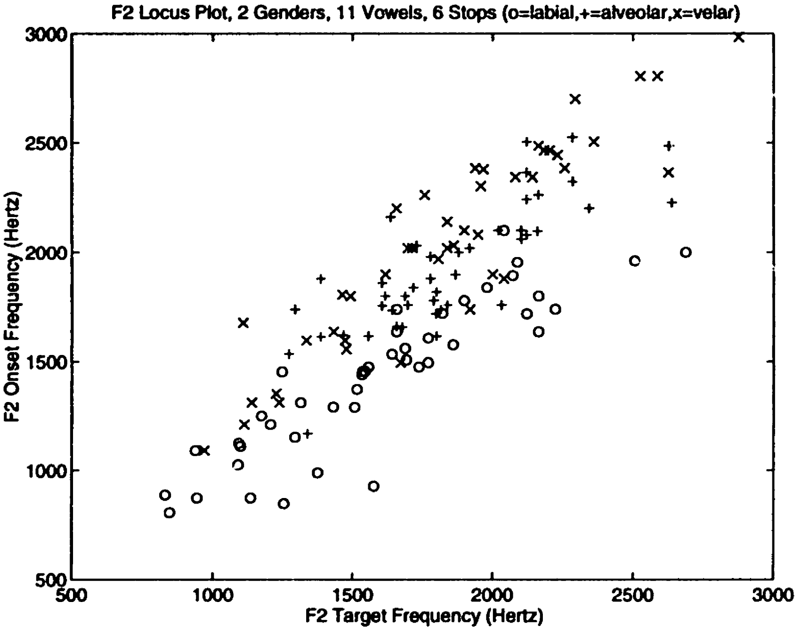


Figure 1-1: F2 onset (first measurable F2) as a function of F2 target (40ms after consonant release), measured by human judges in 131 stop tokens from TIMIT.

phonetic classification is not the primary goal of this thesis. Rather, this thesis seeks to demonstrate that acoustic variability, acoustic measurement error, and phoneme classification error can all be interpreted as the product of articulatory variability, and that the most useful analysis of any one of these manifestations of speech variability is often an analysis which considers all four.

This thesis seeks to develop methods of analysis which draw on the accumulated knowledge of both acoustic phonetics and speech recognition. First, a series of speech production models are drawn from the field of acoustic phonetics, which describe in detail the articulatory sources of all of the types of acoustic variability considered later in the thesis. Second, the field of speech recognition contributes several useful statistical models, including mixture Gaussian error models and a hidden Markov formant tracking algorithm. Finally, the production and the statistical models are used jointly, together with a few results from speech perception studies, to study and interpret the results of three phoneme classification experiments.

1.1 Measurements Used in this Thesis

Before we try to model measurement error, we first must decide what to measure. Following the philosophy set out above, we will try to measure formant and front cavity resonance information at consonant releases. This section will explore what that might mean in a little more detail.

Sussman et al. (1991) have demonstrated discrimination of consonant place on the basis of "locus equations," that is, by modeling the onset frequency of the second formant (F2) as a linear function of its frequency at the center of the vowel. Figure 1-1 is a plot of onset F2 versus vowel center F2, measured by human judges on 66 voiced and 65 unvoiced stop-vowel syllables from the TIMIT database (Zue et al., 1990). (The labeling of these 131 syllables

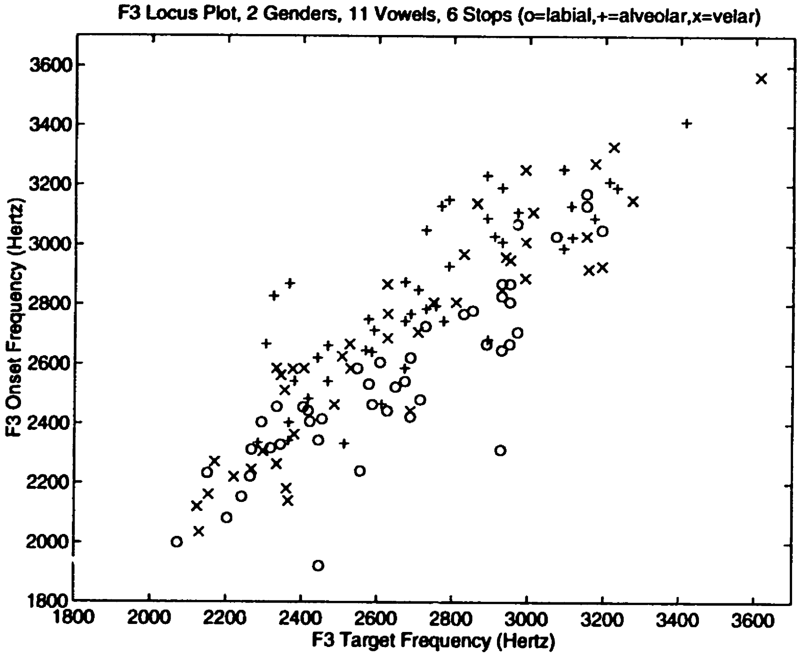


Figure 1-2: F3 onset (first measurable F3) as a function of F3 target (40ms after consonant release), measured by human judges in 131 stop tokens from the "test" subdirectory of TIMIT.

will be described in more detail in chapter 3 of this thesis). This plot is not nearly as clean as the plots in Sussman's study, but it includes a lot of variation specifically excluded in that article: these tokens are extracted from a variety of word contexts, and are produced by more than 100 different speakers, with no preliminary averaging of tokens. Despite the extra variability, there is potentially useful information in this plot. Labial and velar onset frequencies seem to rise in parallel as a function of the vowel target, while alveolar onsets are entirely confined between 1500 and 2500 Hertz (except one /duh/ [1] with an onset at 1200 Hertz). Labial onsets, in fact, are mostly separate from the other two clouds, with the lowest onset frequencies for almost every possible target.

Figure 1-2 is a plot of F3 onset frequency versus vowel center F3, for the same 131 stop releases. The three places of articulation are not as well separated in this plot as they were in figure 1-1, but there is at least a tendency for the F3 onset of labial stops to be lower, for each F3 target, than the onset frequencies of alveolar and velar stops. Perhaps if we combine information from figures 1-1 and 1-2, we can do a little better at separating out the labial stops.

Figure 1-3 displays information from the two previous plots, compressed into two dimensions using linear discriminant analysis. Linear discriminant analysis (LDA) is generally considered a bad algorithm for phonetic classification, because it ignores all information except the means of each cloud, and the total average covariance. One of the main points of Sussman's study, in particular, was that inter-phoneme differences in covariance, which are not modeled by LDA, are important in classifying stops. Despite the obvious drawbacks of LDA, it will be used extensively in this thesis for simple, preliminary classification tasks.

[1] Phonetic quality is represented, in this thesis, using TIMIT notation. For example, /d/ is the voiced alveolar stop release, and /uh/ is the lax high back vowel.

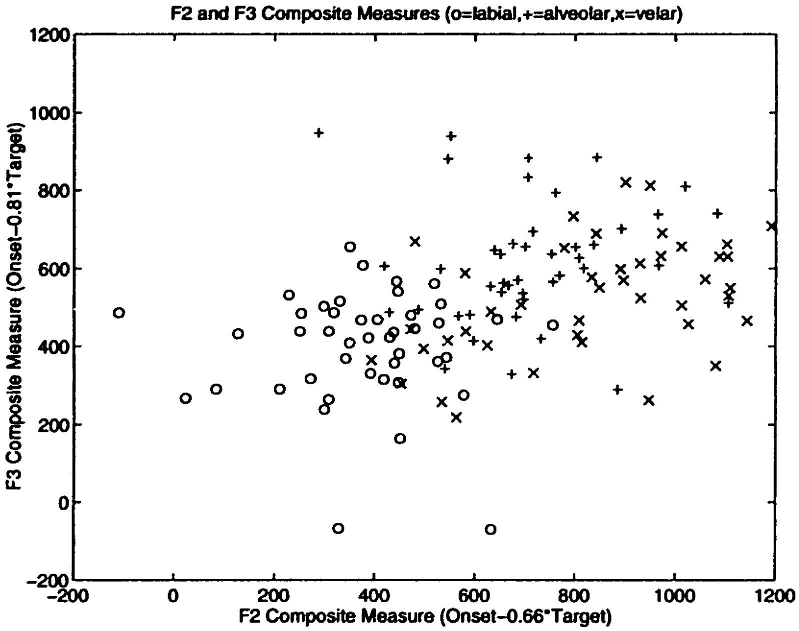


Figure 1-3: Composite F2 and F3 measurements, designed to differentiate lip and tongue places of articulation.

Reasons for this choice will be discussed in section 1.5.

In figure 1-3, the abscissa is an LDA composite of F2 measures, and the ordinate is an LDA composite of F3 measures. Both composite measurements were designed to separate the data into two clouds, one cloud consisting of labial stops, the other consisting of alveolar and velar stops (which we can collectively refer to as "lingual" stops). In this figure, by combining the F2 and F3 locus plots, we have vastly improved the separation of labial and lingual stops. In fact, figure 1-3 may represent the best labial/lingual separation we can get (on this data set) using measurements of F2 and F3. It is time to add another measurement.

Blumstein and Stevens (1979) suggest a classification scheme which takes advantage of the fact that alveolar and velar stops are released into a resonant front cavity, while the turbulent burst of a labial stop is released directly into open space. The resonant shaping of an alveolar or velar stop can add 10 to 20 dB to its amplitude, making alveolar and velar stops typically more intense than labials. Figure 1-4 plots the amplitude of the largest peak in the F2-F6 range of the burst spectrum, as measured by human judges on the same 131 stops, against an LDA composite of the four formant measurements introduced previously. Amplitudes are measured in decibels, with reference to an arbitrary constant. TIMIT recording levels are normalized (Zue et al., 1990); observations by the judges suggest that variation in the recording level probably accounts for no more than 6-10dB of the range of variability in each cloud.

The separation in figure 1-4 is quite good. A reasonable curved boundary between the labial and lingual clouds would result in six or seven errors, or an error rate of about 5%. A linear boundary, calculated, for example, using LDA, would result in at least eight errors.

Now that we've identified most of the labial stops, we need measurements to separate the velar and alveolar stops. Figure 1-5 shows the front cavity resonance, at release of the alveolar and velar stops from the previous figures, plotted as a function of F2 onset. The separation is almost total: the front cavity resonance of velar stops only rises above 2500

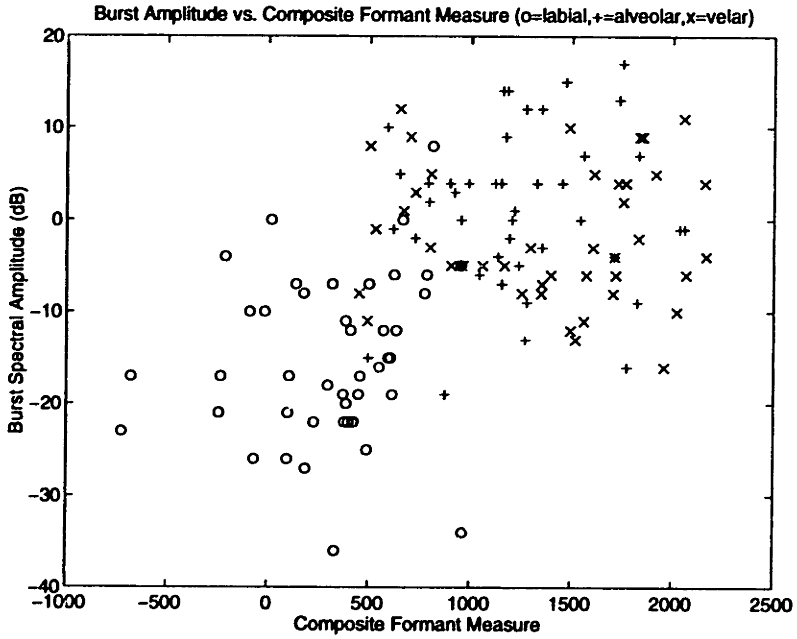


Figure 1-4: Amplitude of the highest peak in the F2-F6 range of the burst spectrum, plotted against a composite formant measurement designed to separate lip and tongue stops. 131 tokens, measured by human judges, from the "test" subdirectory of TIMIT.

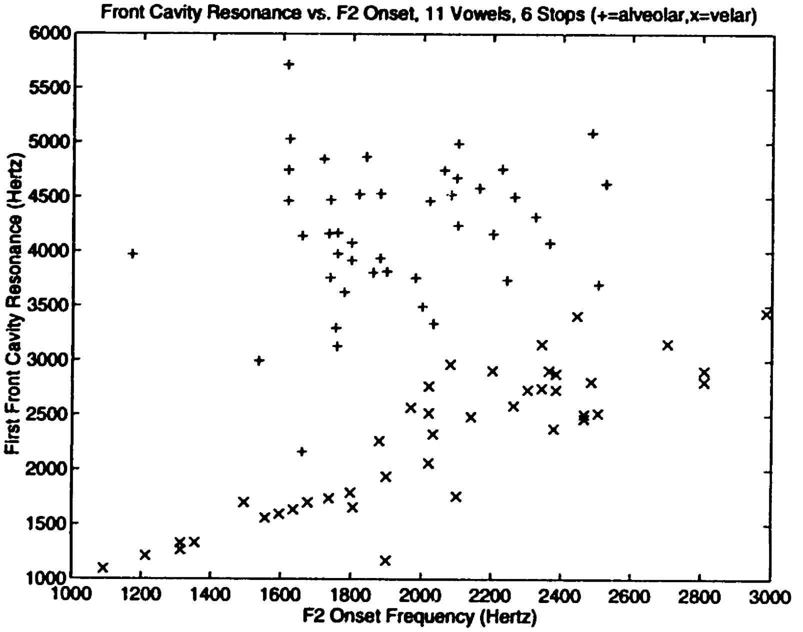


Figure 1-5: Frequency of the front cavity resonance, as a function of F2 onset. 88 alveolar and velar stop releases, extracted from the "test" subdirectory of TIMIT, measured by human judges.

Hertz in front vowel context, that is, when the F2 onset is high, and there is only one alveolar token with a front cavity resonance below 3000 Hertz. In fact, the phonemically alveolar token at 2000 Hertz is not phonetically alveolar: the preceding vowel, an /er/, has pulled the tongue blade back to a retroflex place of articulation. A linear boundary between the alveolar and velar stops in this figure results in complete separation of alveolar and velar tokens; the retroflex alveolar should perhaps receive special handling, because of the novel place of articulation. [2]

To summarize: with only four formant measurements and two burst spectral measurements, we have succeeded in separating all but 6 or 7 out of 131 manually labeled utterances. As a classification argument, of course, this reasoning is somewhat circular. First, we are testing on the training data. Chapter 5 will demonstrate that a simple pairwise LDA classifier, trained on a separate training set, classifies this data with 89% accuracy.

Second, the argument in this section is somewhat circular because the human judges, who had access to the orthography, may have introduced some bias into the measurements. The question of measurement bias is an important one, which is difficult to avoid when asking human transcribers to make complex measurements. The only way to directly address the question of measurement bias is by coding the measurement procedures into an algorithm, and testing for differences between the algorithm and the measurements it is supposed to imitate. If human measurements can be used to classify speech, and if algorithms can imitate human measurements, then it should be possible to classify speech automatically using the algorithms. The problem of classifying speech using human measurements has been addressed here briefly; the problem of writing algorithms to imitate human measurements will be the subject of this thesis.

1.2 Previous Studies: Automatic Formant Tracking

Many high quality automatic formant trackers have been developed over the years. This section will discuss only the two which have contributed most directly to the work in this thesis: an LPC-based tracker, which was used to provide inputs for the rule-based system described in chapter 3 of this thesis, and an earlier hidden Markov model (HMM) tracker, the design of which provided much of the foundation for the tracker developed in chapter 4. This section will discuss these two formant trackers in chronological order, beginning with the HMM tracker.

1.2.1 An HMM formant model

One of the first formant trackers to make use of a global dynamic programming search algorithm was the HMM formant tracker developed by Kopec (1986).

In Kopec's HMM formant tracker, the formant frequencies are viewed as hidden "states" of the speech production mechanism, which condition the production of output spectra. The search space consists of the set of possible formant frequency combinations, simplified by either vector or scalar quantization, in order to reduce the computational load. The transition probabilities are trained on data, and are generally monotonically decreasing as a function of the absolute change in frequency.

[2] Stops with retroflex and lateral right contexts were omitted from this database, specifically to avoid confounding alveolar, retroflex, and lateral places of articulation. The retroflex stop shown in figure 1-5 assimilated retroflex articulation from its left context, which was not examined prior to analysis.

Kopec's tracker uses a discrete output model. The spectrum of a speech frame being examined is first vector quantized using an LPC distance metric, and the local probability of affiliation with each candidate formant is determined by table lookup. These local output probabilities are then combined with the transition probabilities using the forward-backward algorithm, and the resulting *a posteriori* formant probability distribution is used to calculate the conditional expected value of the formant.

The formant frequencies generated by Kopec's algorithm were tested as additional features in an HMM digit recognizer with explicit duration models, whose other acoustic features included the LPC spectrum, and the low-pass and total spectral energies (Bush and Kopec, 1987). As with all previous attempts to use formants in speech recognition, there was no attempt to explicitly model the measurement uncertainty of the formants, and the recognition algorithm may therefore have suffered from formant tracking errors. Since the recognition algorithm was given the same LPC spectra from which the formant frequencies were calculated, the authors argue, the possibly erroneous formant frequencies were effectively a corrupted version of information already available to the recognizer. Even so, the first derivative of F2 was found to increase digit recognition scores. First derivatives of F1 and F3 decreased recognition scores very slightly, while static information about any formant significantly decreased recognition scores.

1.2.2 Finding the roots of the LPC polynomial

Linear predictive coding (LPC) was first proposed as an algorithm for efficiently finding the resonant frequencies of the vocal tract (Atal and Hanauer, 1971), and is still used for that purpose. The formant tracker developed by Talkin (1987) and distributed by Entropic Signal Processing (1993) consists of two stages: a signal processing stage, which generates formant candidates by solving for the roots of an LPC polynomial, and a dynamic programming stage, which finds the most likely sequence of formants from the set proposed by LPC.

In his 1987 presentation, Talkin suggests several technical guidelines for LPC analysis. Perhaps the most interesting of these is his discussion of window choice. Talkin suggests that a window with high spectral sidelobes can significantly degrade a spectrum with a large dynamic range. The formant tracker distributed by Entropics uses, by default, a 49ms cosine-to-the-fourth (cos**4) window. This window has a temporal and frequency resolution similar to that of a 16ms rectangular window or 32ms Hanning window, but has much lower spectral sidelobes than either.

In the second stage of the formant tracking algorithm, all possible mappings between the LPC roots and the desired formants are enumerated, and a Viterbi algorithm is used to find the optimum alignment. Local formant assignment costs are proportional to the bandwidth of the LPC root, and to the absolute deviation between the root frequency and an average formant value. Transition costs are proportional to the change in formant frequency, divided by an estimate of overall spectral change.

1.3 Previous Studies: Acoustic Cues for Place Classification

In deciding what to measure, this thesis will depend on several previous acoustic phonetic studies which have explored the acoustic correlates of consonant place. Most of the acoustic correlates of stop consonant place proposed in the literature fall into three categories: descriptions of the spectral shape of the onset, measurements of formant motion, and time-frequency spectral plots, commonly referred to as dynamic spectra. Of the various acoustic correlates of place proposed for nasal consonants, formant and dynamic spectral measurements are the only acoustic features which have been proven to be useful for classification.

1.3.1 Onset spectral measurements for classification

O.ıe of the first studies investigating the use of onset spectra to classify the place of stop consonants was published by Halle, Hughes, and Radley in 1957. In this study, the authors gated twenty milliseconds from the releases of 99 stops (11 contexts, 3 places of articulation, 3 speakers). The authors noted first that velar stops in front vowel contexts were "acute," with strong peaks above 2000 Hertz, while velar stops in back vowel context were "grave," with spectral peaks at much lower frequencies. The authors therefore adopted a two-tiered classification system. A stop was first classified as either acute or grave, on the basis of a ratio of high-frequency to total spectral energy. Acute stops were then judged to be velar if most of the energy was concentrated between 2000 and 4000 Hertz, and alveolar otherwise. Grave stops were classified as labial or velar based on measurements of the peak frequency, and of the dominance of the largest spectral peak. These measures resulted in about 79% correct classification of place.

Blumstein and Stevens (1979) classified the onset spectra of stops (band-limited to 5000 Hertz) by fitting them to fixed templates. The templates were developed in part based on speech production theory, and in part based on the analysis of 30 training tokens produced by two speakers. The templates were then tested using pre-emphasized LPC spectra with a 26 millisecond window, band-limited at 5000 Hertz, from the onsets of 450 stops uttered by six speakers. The labial template was characterized as "diffuse falling," and required the largest high frequency peak and the largest low frequency peak to fall within 10dB of each other, with variability allowed below about 1500 Hertz. The alveolar template was characterized as "diffuse rising," and required the largest two spectral peaks, with the exception of a possible F2 peak near 1800 Hertz, to fit within a pair of reference lines separated by 10dB and rising at about 10dB per kiloHertz. Finally, the velar template was characterized as "compact," and effectively required a single peak between 1200 and 3500 Hertz to be about 10dB larger than all other peaks in this range. Using these templates, the authors were able to classify the test tokens with about 85% accuracy.

1.3.2 Formant frequency information

It has been known, since the perceptual work of Delattre, Liberman, and Cooper in 1955, that formant frequencies may signal the place of a stop, but the degree to which formant frequencies are actually used for this purpose in naturally occurring consonants has been the subject of considerable controversy.

In 1961, Lehiste and Peterson measured formant transitions at the release of 1263 consonants (one speaker), and found so much overlap between the different consonants that they declared formants to be useless for the classification of naturally occurring stops. Their conclusion was quantified in 1982 by Kewley-Port, who measured, by hand, three-point approximations to the formant trajectories in 120 test tokens uttered by a single speaker. She found that the onset frequencies of F2 and F3, taken together, correctly classified 97% of the tokens given a known vowel context (and given a classifier trained on the test data), but stop place identification independent of vowel context was only 68% correct.

Sussman, McCaffrey, and Matthews (1991) modeled vowel-dependent formant coarticulation as a simple linear relationship between the formant at voice onset and the formant at vowel center. They found that the degree of coarticulation – the correlation between these two measures – varied significantly depending on the place of articulation, and that correct modeling of this correlation was essential to correct classification. They used a quadratic discriminant to classify the average of five repetitions of each of 600 test syllables (3 stops, 10 vowels, 20 speakers), and achieved 77% correct classification of these average utterances.

1.3.3 Dynamic spectral information

Most recent classification studies have used dynamic spectral information as input, typically computed as a simplification of periodic spectra in the first 40 to 70 milliseconds after consonant release. Kewley-Port (1983) classified the three voiced stops with 88% accuracy by using a template method, based partially on the templates of Blumstein and Stevens, with 40 millisecond sequences of LPC spectra (step 5ms) as input. Waibel et al. (1989) used highly-trained speaker-dependent neural networks (three speakers; training and test sets each contained 2620 tokens per speaker), with 150 millisecond mel-frequency spectrograms as input (step 10ms), to classify the three voiced stops with 98.5% accuracy.

In a study comparing several acoustic feature sets, Nossair and Zahorian (1991) realized their best classification performance using a 60 millisecond smoothed cepstrogram (step 10ms) on training and test databases which each consisted of 1260 tokens from 15 speakers (5 male, 5 female, 5 children). The first seven cepstral coefficients (including the spectral mean) were temporally smoothed using a third-order discrete cosine representation, and then modeled using a Gaussian classifier. The six English stops were classified with 94% accuracy; place classification given the voicing of a stop was also roughly 94% correct.

Nossair and Zahorian compared their results with a dynamic spectral representation, described above, to the results achieved with a static onset cepstrum, and with formant frequency and amplitude tracks. The static onset cepstrum was calculated with a 26ms window (meant to imitate the window used by Blumstein and Stevens) and bilinear frequency warping, and yielded 82% correct place classification. Formant frequencies and amplitudes were calculated using an automatic formant tracker; the first three formant frequencies alone yielded 65-70% correct classification, while the combination of formant frequencies and amplitudes yielded 80-85% correct classification.

1.3.4 Classification using TIMIT

All of the published studies reviewed above report classification scores using isolated or stressed monosyllables. In contrast, the speech heard by humans in normal conversation contains stop releases in a wide variety of phonetic and prosodic contexts, and this added variability presumably makes identification of the place of a stop somewhat more difficult.

All of the experiments reported in this thesis rely on data from the TIMIT database (Zue et al., 1990), a national standard for the development of phonetic classifiers. TIMIT consists of transcribed sentences read by 630 speakers. Stop release tokens in TIMIT occur in a wide variety of phonetic and prosodic contexts.

One of the first classification studies using data extracted from TIMIT was published by Lamel (1988), who classified stop consonants using a rule-based classification system. Based on manual transcriptions of formant frequency, burst information, and voice onset time, Lamel reported 90% correct classification of stop consonant place.

Most state of the art classifiers depend on a sort of dynamic spectral representation, usually consisting of mel frequency cepstral coefficients (MFCCs) and their temporal deriva-

Spoken	Number	Classified As:		
Place	of Tokens	labial	alveolar	velar
labial	530	90%	6	4
alveolar	652	3	91	6
velar	503	2	6	92

Table 1.1: Stop place classification scores derived from the data of Chun (1996). The classifier is trained using the TIMIT TRAIN database (3696 sentences), and classification scores are reported for the TIMJT DEV database (400 sentences).

tives. Chun (1996) reported classification results using 36 time-averaged MFCCs (12 coefficients × 3 frames), 24 MFCC derivatives (12 × 2), and the logarithm of segment duration. Results for the classification of place of unvoiced and voiced stops have been extracted from his data, and are shown in table 1.1; overall classification of stop place is 91% correct.

1.4 Discussion: Acoustic Correlates of Place

The available studies show remarkable agreement on the sufficiency for place classification of the three reviewed types of acoustic correlates. Burst spectral cues are apparently sufficient to classify place with 80-85% accuracy. Formant frequencies, taken alone, are sufficient to classify place with about 65-70% accuracy (Sussman et al. apparently achieved higher rates by averaging out some of the relevant variability). Dynamic spectral cues are sufficient to classify place with greater than 90% accuracy across speakers, and with 98.5% accuracy using Waibel's highly trained speaker-dependent model.

This thesis proposes to study formant frequencies and burst spectral cues for use in the classification of consonant place. Our experiments in section 1.1, and in chapter 5, support the conclusions of Lamel (1988), who found that a combination of formant frequencies and burst spectral information measured by human judges can be used to classify the place of TIMIT stop consonants with about 90% accuracy. Since the best reported classification of stop place in TIMIT is about 91% correct (Chun, 1996), we can conclude that burst spectral measurements and formant frequencies measured by human judges are sufficient to classify stop place with an accuracy roughly equal to the state of the art.

Duplicating the measurements of human judges automatically, however, is difficult. Automatic measurements of formant frequencies and burst spectral measurements do not seem to have been used together for classification in the past, but Nossair and Zahorian's (1991) automatic measurements of formant frequency and amplitude provided significantly worse classification than a periodic cepstral measurement. Formant and burst spectral measurements, it seems, suffer degradation caused by acoustic measurement error, while a dynamic cepstral representation does not. If the primary goal of this thesis were classification, therefore, it would be logical to begin with a state of the art dynamic cepstral representation.

In this thesis, phonetic classification is not the primary goal. Instead, classification is used here as a tool for the analysis of speech production variability, and the primary goal is an integrated analysis of variability in production, acoustics, and phonetic classification. The goal of an integrated analysis is most easily reached if the acoustic measurements reflect known relationships between articulation and acoustics.

Formant frequencies, and the front cavity resonance of a burst spectrum, can be defined in either articulatory or acoustic terms. In this thesis, both measurements are defined in articulatory terms (as the eigenfrequencies of the vocal tract, and of specified vocal tract cavities, respectively), but the acoustic definition (as a particular set of ordered spectral peaks) is closely related, and often results in almost identical parameter values. Since formant frequencies and the front cavity resonance can be defined in either articulatory or acoustic terms, they are not strictly either articulatory or acoustic measurements. Rather, the measurements of formants and front cavity resonance at a particular stop release can be defined as a physical instantiation of aspects of the theory of acoustic speech production, effectively serving as a bridge between uniquely articulatory and uniquely acoustic measures.

As a bridge between articulation and acoustics, formant frequencies and the burst front cavity resonance are uniquely suited for the integrated analysis of variability proposed in this thesis. All error analysis and classification experiments in this thesis are therefore based on acoustic measurements of formant frequency and burst spectral characteristics.

1.5 Discussion: Classification as a Tool for the Analysis of Variability

The goal of this thesis is an analysis of acoustic speech variability in terms of the known relationships between articulation and acoustics. In order to make the results as accessible as possible to other researchers, the experiments in this thesis use data from a large speech database (TIMIT) which is available to all interested researchers from the Linguistic Data Consortium, a broad consortium of companies, universities, and government agencies based at the University of Pennsylvania. The choice of a purely acoustic database, however, limits the degree to which articulatory explanations of variability can be confirmed using statistical methods.

Two types of variability are the focus of most of the analysis in this thesis. First, measurement errors are analyzed extensively: production models are developed to analyze the sources of measurement error, and statistical tools are used to describe the distribution and context dependence of error. Second, variability in formant frequencies and burst spectral measurements is analyzed as a function of consonant place, and of phonetic context.

Acoustic measurement variability as a function of phonetic category is generally analyzed, in the literature, using one of two types of statistical tool. Analysis of variance (and other similar tools) seeks to determine whether the difference between categories, compared to the variation within a category, is too large to have been produced by random variation. Classification analysis, on the other hand, seeks to determine whether there is any overlap between categories, that is, whether the difference between categories is sufficiently large to completely separate the categories.

Variation of acoustic measures as a function of consonant place, and the interaction between measurement error and consonant place, are primarily analyzed in this thesis using the tools of phonetic classification. Classification analysis is chosen as a more useful tool than analysis of variance for three reasons.

First, classification analysis is, in a sense, less forgiving than analysis of variance. Analysis of variance seeks only to find out whether the phonemes are separated; classification analysis attempts to find measurements which separate them with no overlap.

Second, classification analysis using formant and burst spectral measurements can be easily compared to results published in the literature. Thus, for example, the effect of measurement error is characterized in chapter 5 of this thesis as a drop in a phonetic classification score, from 89% using manual measurements, to 76% using automatic measurements. Both numbers can be compared, with reasonable confidence that the comparison is warranted, to the classification scores obtained by other researchers using the TIMIT database.

Third, the separation of consonant place using classification analysis can be compared to the ability of human listeners to discriminate stops. Variability in the acoustic production of speech sounds is easily measured, but it is often more difficult to characterize the relationship between acoustic variability and perceptual errors. Chapter 5 will compare the error pattern of a phonetic classifier, as a function of phonetic context, to the error patterns of human listeners. This analysis is intended to suggest that the sources of perceptual error may be modeled by analyzing the sources of classification error, although any more rigorous pursuit of this suggestion is beyond the scope of this thesis.

For the reasons given above, phonetic classification is used extensively in this thesis as a tool for the analysis of speech production variability. Before proceeding, however, we still need to discuss the choice of a classifier structure.

In this thesis, acoustic measurement error and phonetic classification error are treated as manifestations of acoustic variability. It turns out that the relationship between classification error and acoustic variability can be modeled very precisely by feeding a parametric model of the acoustic measurement distribution to the classification rules of a parametric classifier. This method for modeling classification error will be described in section 5.1.

Of the available parametric classifiers, a classifier based on linear discriminant analysis (LDA) is perhaps the easiest to visualize, and to analyze numerically. In an LDA classifier, a set of acoustic measurements is first weighted and summed to produce a one or two dimensional linear discriminant representation, and then the linear discriminant representation is classified using a fixed classification threshold. The linear discriminant representation of the data can be plotted for visual inspection (as in section 1.1), and the probability of a measurement crossing the threshold and being misclassified can be expressed in closed form as the integral of the acoustic measurement distribution (as shown in section 5.1).

Linear discriminant analysis is not the best structure for phonetic classification, as discussed in section 1.1, but it lends itself well to an analysis of classification error as a function of acoustic variability. In this thesis, classification is a tool, used in the pursuit of better models of variability. Since LDA classification lends itself to the more important analysis goals, it is used in most of the reported classification experiments in this thesis.

1.6 Thesis Outline

The goal of this thesis is an integrated analysis of speech production variability, in which the descriptive power of statistical models is combined with the explanatory power of speech production models. The organization of the thesis is designed to build the reader's understanding of speech variability in three stages. First, chapter 2 describes variability in models of speech production, and the link between articulatory and acoustic variability. Second, chapters 3 and 4 describe the link between acoustic variability and acoustic measurement error. Finally, chapter 5 describes the dependence of classification error on acoustic measurement error, and chapters 5 and 6 discuss the extent to which the relationship between acoustic variability and phonetic classification can be used as a model of the relationship between variability and human speech perception.

Chapter 2 demonstrates that models of speech production can be used to explain, and under certain circumstances to predict, characteristics of acoustic variability. A model of turbulence noise is developed to show that in some cases, the form of variability in the speech spectrum can be derived entirely from physical principles, without considering variations in speaker anatomy or speaking style. The more common case, in which anatomy and speaking style play an important role in the form of variability, is exemplified by a discussion of the transfer function amplitudes of front and back cavity resonances during frication. In this case, it is argued that physical principles can set useful limits on the range of variability, but that more detailed knowledge of the form of variability must be gathered empirically. A method for the empirical study of variability is proposed, in which variability in acoustic phonetic measurements is described using statistical models, and explained using speech production models.

Chapters 3 and 4 describe procedures for combining production knowledge and empirical observation in the design of acoustic measurement algorithms. In chapter 3, a knowledge engineering approach is attempted, in which a trained phonetician designs rule-based algorithms to imitate his own formant and burst spectral measurements on a training corpus. The size and frequency of measurement errors produced by the finished algorithms are then evaluated on an independent test set, and aggregate statistical models of the distribution of error are developed. Finally, the usefulness of aggregate models of measurement error in studies of speech production variability is discussed. It is argued that many acoustic measurement errors can be predicted by the presence of ambiguities in the acoustic spectrum, that this information is useful, and that the aggregate error models developed in chapter 3 ignore this information.

Chapter 4 demonstrates that the uncertainty in a given formant measurement can be predicted from the presence of measurable ambiguities in the acoustic spectrum. A novel procedure for combining production knowledge and empirical observation is suggested, in which production knowledge guides the design of an HMM formant tracking algorithm, which is then trained on empirical data. It is shown that in formant tracking applications, an HMM formant tracker is able to generate *a posteriori* estimates of the measurement uncertainty for each formant, based on the information about acoustic cues contained in the model structure and parameters. The *a posteriori* uncertainty estimates generated by the proposed formant tracker are evaluated by comparison to the known measurements of two human judges on a test set.

Finally, chapter 5 describes several phonetic classification experiments which explore the relationships among acoustic variability, acoustic measurement error, and classification error. First, linear discriminant classification of place is tested using both manual and automatic acoustic measurements. The difference between the performance using manual measurements and the performance using automatic measurements is taken to be the effect of measurement error, and it is shown that the difference in classification performance is well predicted by the aggregate error models of chapter 3. Second, both rule-based and HMM measurement algorithms are used in context-dependent classification of place over a large database. While the total classification score is somewhat disappointing, the pattern of errors as a function of phonetic context is shown to be similar to the pattern of errors of human listeners, indicating that the kinds of acoustic variability which most confuse the classifiers may be the same kinds of variability which confuse human listeners.

Chapter 6 summarizes conclusions, and suggests future work.

Chapter 2

Speech Production Theory

The speech production mechanism can be modeled as the filtering of a variety of high-impedance, nonlinear sources through two linear filters (Fant, 1960). The first filter, $T(f)$, represents the vocal tract transfer function from a source flow $S(f)$ to the volume velocity at the mouth,

$$U_m(f) = T(f)S(f) \tag{2.1}$$

The second filter is the radiation characteristic $R(f)$, which models the conversion from mouth flow to radiated sound pressure,

$$P_r(f) = R(f)U_m(f) \tag{2.2}$$

In this thesis, the radiation characteristic is assumed, throughout, to be the characteristic of a simple point source,

$$R(f) = j\frac{\rho f}{2r}e^{-j2\pi fr/c} \tag{2.3}$$

where ρ is the density of air (approximately 0.00112g/cm^3), and r is the distance between the mouth and the microphone.

This chapter covers variability in the source spectrum $S(f)$, variability in the transfer function $T(f)$, and finally, predicted correlations between $T(f)$ and the place of articulation of a consonant. The first two sections each conclude with a discussion of measurement issues stemming from, respectively, source variability and transfer function variability; when work in these sections is believed to be original, it is identified as such. Readers who are not already familiar with speech production theory may find this review overly concise, and may wish to refer to the more thorough presentations in Flanagan (1972) or Stevens (in preparation).

Although this chapter is intended to serve as a review of speech production theory, it is also intended to contribute to our understanding of speech variability. This chapter begins by reviewing standard acoustic phonetic models of average speech production, of the type which are often used in speech analysis and speech synthesis. After establishing models of average production, however, some of the sections in this chapter proceed to develop acoustic phonetic models of the range, or even of the probability distribution, of speech production variability. Some of these models will be used in chapter 3 as background for the design of acoustic correlate measurements, but all of them are also intended to serve as examples in support of one of the hypotheses central to this thesis. The discussions of variability in this chapter are intended to support the hypothesis that speech production

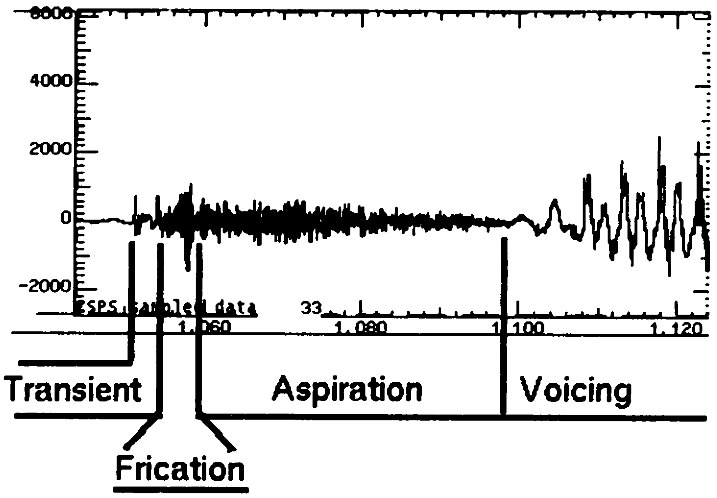


Figure 2-1: During the release of an unvoiced stop, four distinct acoustic sources are activated. The source contributions overlap in time; for example, ringing of the transient continues past the onset of frication.

models, which have been used successfully in the past to explain sample acoustic correlate measurements, are also helpful in explaining the aggregate distribution of acoustic correlate variability.

2.1 Speech Sources

During the release of a stop, three or four distinct acoustic sources are activated, in the sequence shown in figure 2-1, with some temporal overlap between sources. First, when the oral constriction is released, the pressure drop across the constriction is equalized with an audible air flow transient, which may excite the transfer function with sufficient strength to ring audibly for several milliseconds. During and after the ringing of the transient, turbulent flow develops in the constriction, generating frication sources at the constriction for 5-20 milliseconds or more.

As pressure drops in the vocal tract, frication ceases, and the source of excitation shifts to the glottis. If the stop is unvoiced and syllable initial, the vocal folds are actively held open for 40-100 milliseconds after release, during which time the turbulent glottal jet generates audible aspiration noise. If the stop is voiced, on the other hand, regular sonorant voicing begins as soon as the oral pressure is low enough, and usually within 25 milliseconds after release.

This section reviews the spectral shapes of transient, turbulent, and voicing sources.

2.1.1 Transient source

When the oral constriction is first opened during release of a stop, the pressure drop across the constriction is equalized with an audible air flow transient.

Given an adequate model of the variable resistor in figure 2-2, it is possible to apply standard transmission line theory to predict the shape of the acoustic transient. The relationship between pressure and flow across most vocal tract constrictions can be ap-

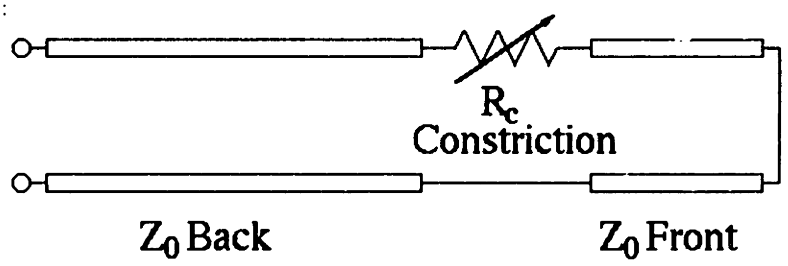


Figure 2-2: Transmission line model of the vocal tract configuration immediately after release of a stop. The characteristic impedance of the vocal tract is Z_0, the constriction resistance is $R_c(t)$, the glottal impedance is set to infinity, and the radiation impedance is set to zero. The back cavity is charged to a non-zero pressure, which is equalized quickly as the admittance of the constriction grows.

proximated using a conservation of energy constraint, as follows. Assuming that the area A_t upstream from the vocal tract is much larger than the area A_c of the constriction, the average velocity of individual air particles v needs to increase considerably upon entering the constriction in order to maintain a constant volume velocity $U_0 = vA$. By equating the increase in kinetic energy to a corresponding loss in potential energy, we derive Bernoulli's equation (Flanagan, 1972):

$$P_0 = \frac{k\rho}{2}\left(\frac{U_0}{A_c}\right)^2 \tag{2.4}$$

where $k \approx 1$ is a constant which depends on the shape of the constriction. The equivalent acoustic resistance of the constriction can be derived by linearizing equation 2.4 for small-scale perturbations to a relatively steady-state gross flow, $U = U_0 + dU$:

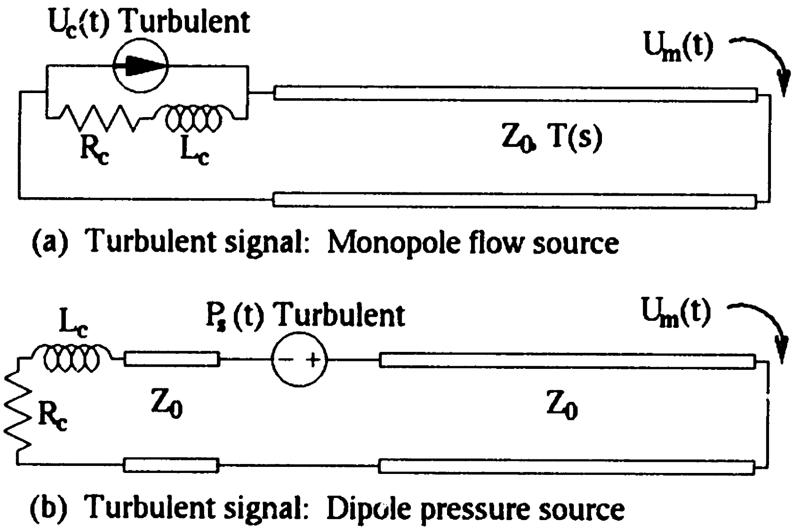
$$P = P_0 + R_c dU, \quad R_c = \frac{k\rho U_0}{A_c^2} \tag{2.5}$$

Based on equation 2.4 and standard transmission line theory, Massey (1994) has shown that the flow through the constriction initially grows in direct proportion to the constriction area. Assuming that the area of the constriction at the moment of release is well approximated by some power of t, $A_c(t) \approx a_0 t^\alpha$, the initial flow transient spectrum (before consideration of the radiation characteristic) is proportional to $a_0 f^{-(\alpha+1)}$ at higher frequencies. Empirically, Massey has shown that the transient spectrum typically shows an f^{-2} dependence.

2.1.2 Turbulent sources

Stevens (1971) has described two types of sources which are important in the production of what we perceive as turbulence noise. Figure 2-3 shows schematized circuit models for these two sources.

The first source, called a "monopole source," consists of random fluctuations in the flow through the constriction. These random fluctuations can be modeled as a flow source in parallel with the constriction impedance, as shown in figure 2-3a. The spectrum of the flow source has a low-pass characteristic, as shown in figure 2-4a. The amplitude of this source may vary considerably, depending on the length of the constriction, and depending on whether or not there are any flow obstacles upstream from the constriction.



(a) Turbulent signal: Monopole flow source

(b) Turbulent signal: Dipole pressure source

Figure 2-3: Circuit models for the monopole and dipole turbulent acoustic sources. Radiation impedance and impedance of the back cavity are ignored; the mouth flow $U_m(t)$ is just the flow through a short circuit at one end of the transmission line. The monopole source $U_c(t)$ is in parallel with the constriction impedance; the dipole source $P_s(t)$ is connected in series, typically 1-3cm downstream from the constriction.

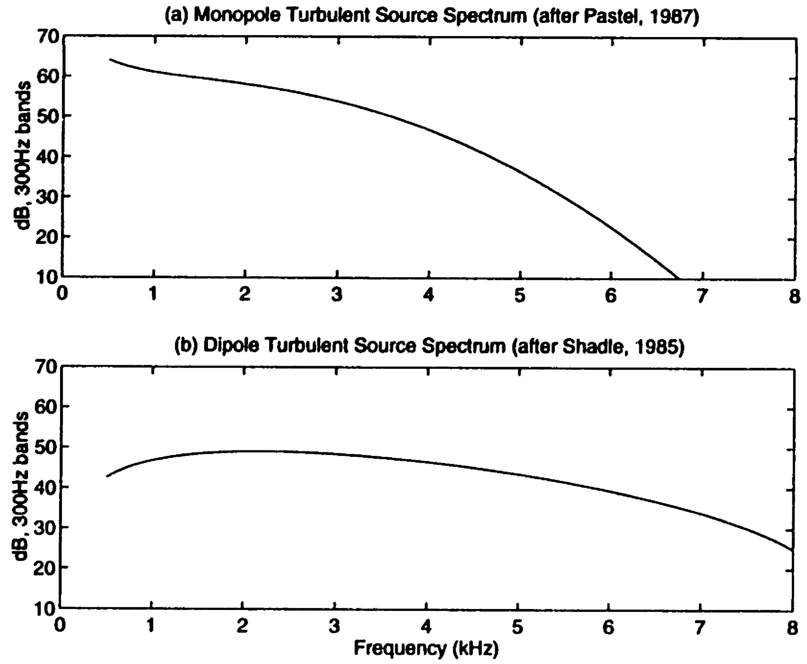


Figure 2-4: (a) Monopole turbulent source spectrum, based on the radiated spectra reported by Pastel (1987). (b) Flow source spectrum composed of the product of a dipole turbulent source, based on spectra reported by Shadle (1985), and a coupling factor $G_c(f)$ which assumes a vocal tract area of 5cm^2 and a distance of 2cm between constriction and source. Amplitudes reported by Pastel and Shadle have been adjusted to represent 300Hz bands, a flow of 420 cm^3/s, and a constriction area of 0.08 cm^2.

The second source, called a "dipole source," is caused by the collision of turbulent vortices with obstacles downstream from the constriction. These collisions cause random pressure fluctuations, which can be modeled as a series pressure source as shown in figure 2-3b. The amplitude of this source depends on several factors, including at least the shape and angle of the flow obstacle (Gordon, 1969). The shape of the obstacle, and therefore the shape of the total turbulent source spectrum, depends significantly on the location and shape of the constriction.

The transfer function between the dipole pressure source, $P_s(f)$, and the mouth flow $U_m(f)$ contains zeros at frequencies approximately equal to $nc/2x$, where x is the distance from the constriction to the flow obstacle, and $n = 0, 1, 2, \ldots$. For most acoustic purposes, these zeros in the transfer function can be combined with the source spectrum to form an equivalent flow source spectrum $U_s(f)$:

$$U_s(f) \equiv G_c(f)P_s(f), \quad G_c(f) = \frac{1}{Z_0} \sin \frac{2\pi f x}{c} \tag{2.6}$$

where Z_0 is the characteristic impedance of the vocal tract. For an obstacle about 2cm downstream from the glottis, the spectrum of the equivalent flow source has roughly the frequency dependence shown in figure 2-4b.

Aspiration

The jet of air coming through the glottis is always turbulent, and therefore always produces turbulent noise. During modal voicing, however, this turbulence noise is usually hidden by the voicing spectrum. Aspiration usually becomes audible when the glottis is actively opened in order to produce an /h/, or the aspirated onset of an unvoiced stop.

In aspiration, the dipole source dominates most of the spectrum. Since the coupling function $G_c(f)$ has a zero at low frequencies, and because the radiation spectrum is proportional to f, aspiration only strongly excites poles with frequencies above about 1000 Hertz. In particular, during the early part of aspiration after a stop release, the first formant is usually below 500 Hertz, and is therefore only weakly excited by the aspiration source. Since the bandwidth of the first formant is also quite wide during aspiration (see section 2.2.2), it is often difficult to find any spectral evidence for the first formant during aspiration.

Frication

When an oral constriction, formed by the tongue or lips, is at least as narrow as the glottal constriction, pressure builds up across the oral constriction, and flow through the constriction becomes turbulent. This frication turbulence generates monopole and dipole acoustic sources similar to the sources produced at the glottis during aspiration.

As with aspiration, the shape of the frication source spectrum depends on the relative contributions of the monopole and dipole sources, and therefore, on the efficiency of the flow obstacles which cause dipole sources. Since the efficiency of the flow obstacle is a function of place of the consonant, it is possible that place of the consonant may be reflected in the shape of the source spectrum. During release of a velar stop, for example, Stevens (in preparation) has calculated that the dipole source dominates the spectrum above 2000-3000 Hertz, and the monopole source dominates the lower-frequency spectrum. During release of an alveolar stop, the dipole spectrum is typically 5-10dB more intense than that of a velar, giving the source spectrum a high-pass characteristic. During release of a labial stop, the

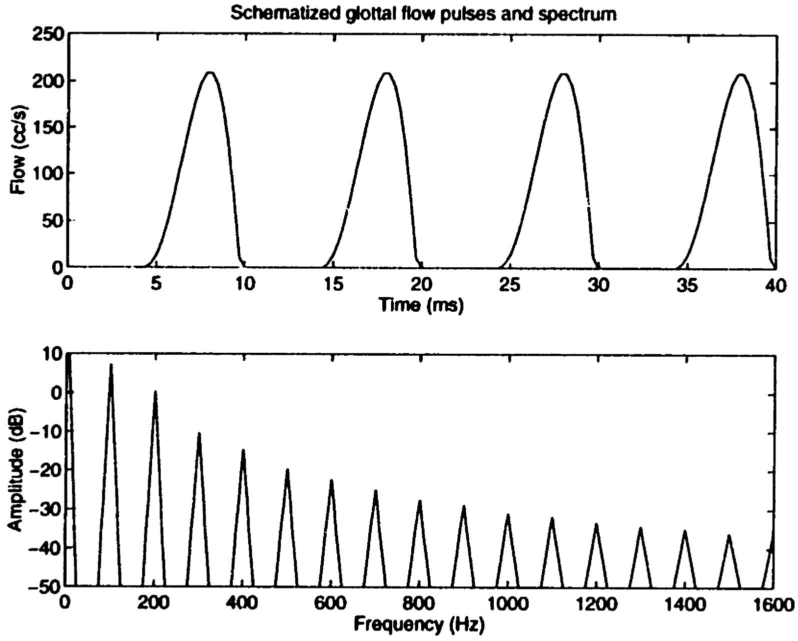


Figure 2-5: Schematized glottal flow waveform and spectrum.

dipole spectrum is typically 5-10dB less intense than the dipole spectrum of a velar, giving the source spectrum a low-pass characteristic. Stevens notes, however, that these relative amplitudes are quite variable, and may depend on both speaker and phonetic context.

2.1.3 Glottal vibration

In normal glottal vibration, the pressure drop between the trachea and the vocal tract drives a jet of air through the glottis, which in turn drives the vocal folds to vibrate, much like a flag flapping in a strong wind (Titze, 1994) . The opening and closing of the glottis modulates the glottal jet into a series of somewhat triangular flow pulses. This pulse train excites the vocal tract transfer function, and a delayed, filtered version of the pulse train is radiated from the mouth.

Figure 2-5 shows a schematized glottal flow waveform and spectrum, based on the parametrized model of Fant, Liljencrants, and Lin (1986). The waveform is periodic with period $T0$, and the spectrum is only non-zero at frequencies which are multiples of $F0 = 1/T0$. The waveform is roughly triangular, and there is a slope discontinuity at the instant of glottal closure. The slope discontinuity in the waveform transforms into a magnitude DFT proportional to $1/f^2$ at high frequencies, equivalent to a log-magnitude DFT with a slope of -12dB/octave.

2.1.4 Measurement issues: time-averaged power spectrum

Glottal vibration and transient sources are deterministic: a model of a voicing or transient source, adequately fitted to the data, tells us exactly the form of the Fourier transform $S(f)$. Given the form of $S(f)$, and a measurement $X(f)$ of the spectrum of speech radiated from the mouth, it is possible to estimate the transfer function (at frequencies where $S(f)$ is non-zero) by just dividing, $T(f) = X(f)/(R(f)S(f))$.

Turbulent sources are stochastic: the Fourier transform $X(f)$ of any finite length turbulent signal is a random vector. A model of a turbulent source can describe the expected value of the source spectrum, and perhaps even the distribution of its components, but no model will ever predict the exact value of an observation of $X(f)$. There is a tendency, in speech analysis, to assume that any observation of the Fourier transform of the radiated speech waveform is approximately equal to its expected value, $E[X(f)] = R(f)T(f)E[S(f)]$. This section sketches an original derivation of the probability density of the squared spectrum $|X(f)|^2$, and demonstrates that $|X(f)|^2$ is often quite different from its expected value. The time-averaged power spectrum (discussed in, for example Shadle 1985) is then presented as a better estimate of the expected value $E[|X(f)|^2]$, and, by way of proof, an original derivation of its probability density is presented.

If a turbulent acoustic signal is modeled as a stationary Gaussian random process, and assuming that the radiation characteristic has removed any mean flow, the acoustic signal $x(t)$ is completely characterized by its power spectrum (Papoulis, 1984), that is, by the Fourier transform of the autocorrelation:

$$P_x(f) = \int_{-\infty}^{\infty} E\left[x(0)x(\tau)\right] e^{-j2\pi f\tau} d\tau \tag{2.7}$$

where the operator $E[]$ denotes expectation, and $E\left[x(0)x(\tau)\right]$ is the autocorrelation of x. Linear system theory tells us that the power spectrum of radiated speech is calculated by multiplying the power spectrum of the source, $P_s(f)$, by the squares of the transfer function and radiation characteristic:

$$P_x(f) = |R(f)|^2|T(f)|^2P_s(f) \tag{2.8}$$

Thus, if we know the power spectrum of the source, and given a good estimate of the radiated power spectrum, the vocal tract transfer function can be estimated by simple division. The shape of the source power spectrum has been described in section 2.1.2. This section considers the problem of accurately estimating the power spectrum of radiated speech.

The power spectrum is often estimated using a magnitude-squared short-time Fourier transform (squared STFT):

$$P_x(f) \approx 2\pi|X(f)|^2, \quad X(f) = \int_{-T/2}^{T/2} w(t)x(t)e^{-j2\pi ft}dt \tag{2.9}$$

where $w(t)$ is some window function which is zero for $|t| > T/2$, and $x(t)$ are samples of the radiated speech signal. The squared STFT is a biased estimator of $P_x(f)$, but it is an unbiased estimator of the smoothed power spectrum

$$2\pi E\left[|X(f)|^2\right] = P_x(f) * W^2(f) \tag{2.10}$$

where $W(f)$ is the transform of $w(t)$, and $*$ indicates convolution.

The real part $Re\{X(f)\}$ of the Fourier transform is just a weighted sum of zero-mean Gaussian random variables, and is therefore itself a zero-mean Gaussian random variable. The square of any zero-mean Gaussian random variable is a scaled first order χ^2 random variable, where the order of the χ^2 variable can be denoted with a subscript:

$$\frac{Re\{X(f)\}^2}{E\left[Re\{X(f)\}^2\right]} = \chi_1^2 \tag{2.11}$$

where the probability distribution of a χ_n^2 variable can be found in, for example, (Drake, 1988). If $Re\{X(f)\}$ and $Im\{X(f)\}$ can be assumed to be have the same variance,

$$E\left[Re\{X(f)\}^2\right] = E\left[Im\{X(f)\}^2\right] = \frac{1}{2}E\left[|X(f)|^2\right] \tag{2.12}$$

then their squares can be added to produce a scaled second order χ^2 random variable:

$$|X(f)|^2 = Re\{X(f)\}^2 + Im\{X(f)\}^2 = \frac{(\chi_1^2+\chi_1^2)}{2}E\left[|X(f)|^2\right] = \frac{\chi_2^2}{2}E\left[|X(f)|^2\right] \tag{2.13}$$

The variance of χ_2^2 is 4, so

$$Var(|X(f)|^2) = \left(E\left[|X(f)|^2\right]\right)^2 \tag{2.14}$$

In other words, the squared STFT is a particularly inefficient estimator of the power spectrum: the ratio of the standard deviation of the estimate to its expected value is 1.0. [1]

The inefficiency of the Fourier transform as an estimator of the power spectrum is well known. Papoulis suggests reducing the estimator variance by smoothing the spectrum. This thesis makes use of a time-averaged power spectrum (see for example Shadle, 1985) which is functionally equivalent to power-spectral smoothing (using appropriate windows), but requires significantly less computation.

A time-averaged power spectrum $\widehat{P_x(f)}$ is the average of several squared STFT spectra, computed using temporally sequential windows. If the signal is considered stationary, and the windows do not overlap, then each squared DFT $|X_n(f)|^2$, $n = 1, \ldots, N$, is an independent, identically distributed estimate of the power spectrum. In particular, if Var($Re\{X\}$)=Var($Im\{X\}$), samples of the average squared spectrum are χ^2 variables of order $2N$:

$$\widehat{P_x(f)} = \frac{1}{N}\sum_{n=1}^{N}|X_n(f)|^2 = \frac{\chi_{2N}^2}{2N}E\left[|X(f)|^2\right] \tag{2.15}$$

By taking logarithms, we can separate $\widehat{P_x(f)}$ into a purely deterministic mean component, and a zero-mean random component:

$$10\log_{10}\widehat{P_x(f)} = 10\log_{10}E\left[|X(f)|^2\right] + 10\log_{10}\frac{\chi_{2N}^2}{2N} \tag{2.16}$$

Based on equation 2.16, it is possible to calculate confidence limits on the amplitudes of spurious peaks and valleys in a time-averaged power spectrum. For example, a spectral estimate composed of the average of two consecutive spectra is distributed as a $\chi_4^2/4$ random

[1] The assumption that Var($Re\{X\}$)=Var($Im\{X\}$) is satisfied by white noise filtered by an LTI system, but may not be satisfied for other stochastic signals. Without using this assumption, Papoulis (p. 494) derived an inequality stating that the standard deviation of $|X(f)|^2$ is greater than or equal to its expected value. Equation 2.13 can be viewed as a special case of his result, in which the more restrictive assumption allows us to derive the probability distribution exactly.

variable. Based on standard χ^2 tables, we can calculate that about one percent of all such spectral samples are more than 11dB below their expected values, and one percent are more than 5.2dB above. Notice the imbalance between positive and negative variation: random spectral nulls are much more likely than random spectral peaks.

If neighboring frequency components are independent (as is true of Gaussian wl te noise processes), a random spectral null or peak in $\widehat{P_x(f)}$ almost always has the shape and width of the transformed window, $W^2(f)$. Using a 6ms Hanning window, for example, any grid of spectral samples separated by frequencies of $2/0.006 \approx 330Hz$ can be viewed as independent. If the spectrum is computed as the average of two consecutive 6ms Hanning windows, the analysis of the previous paragraph suggests that about one out of every hundred 330Hz bands measured in running speech (with non-overlapping windows) contains a randomly generated spectral peak of 5.2dB or more, while an equal number contain a randomly generated spectral null of 11dB or more. The half-power bandwidth of these spectral peaks and nulls is equal to the half-power bandwidth of $W(f)$, which, for a Hanning window, is slightly less than the 330Hz band spacing.

In reality, time-averaged power spectra are almost always computed using overlapping temporal windows. If rectangular windows are used, it is possible to prove linear dependence between STFT transforms computed using overlapping windows, so there is no theoretical advantage to using overlapping windows. If non-rectangular windows are used, the spectral samples computed using overlapping windows are correlated, but not linearly dependent, so using overlapping windows may improve the spectral estimate. Windows with tapered edges, for example, are minimally dependent on waveform samples at the edges of the window; heuristically speaking, therefore, a spectral estimate $\widehat{P_x(f)}$ computed using windows which overlap by about $T/2$ should contain more information about the power spectrum than an estimate computed using non-overlapping windows.

2.2 Speech Filters

This section discusses the relationship between positions of the articulators and the vocal tract transfer function. Factors which may make formants or front cavity resonance peaks difficult to measure, including pole-zero pairs and changes in formant bandwidth, are discussed in some detail.

2.2.1 All-pole models

The vocal tract during a vowel is often modeled as an acoustic transmission line, with no side branches, and with no coupling through the glottis between the vocal tract and the trachea. To the extent that this model is correct, the vocal tract transfer function can be modeled with an all-pole spectrum:

$$T_{ap}(f) = \prod_{n=1}^{\infty} \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \tag{2.17}$$

where $s = j2\pi f$ is the complex radial frequency in Hertz, $s_n = j2\pi Fn - \pi Bn$ is the complex pole frequency, composed of the formant frequency Fn and bandwidth Bn, and s_n^* is the complex conjugate of s_n. For example, if the vocal tract is modeled as a uniform tube of length l, the formant frequencies are quarter-wave resonances of the tube:

$$Fn = \frac{(2n-1)c}{4l} \tag{2.18}$$

During the frication burst of a velar or alveolar stop, the cavity in front of the constriction can be productively modeled as a short uniform tube of length l_f, with no coupling through the constriction to the back cavity. According to this model, the transfer function $T(f)$ is as given in equation 2.17, but with resonant frequencies F_{fn} which are quarter-wave resonances of the front cavity

$$F_{fn} = \frac{(2n-1)c}{4l_f} \tag{2.19}$$

The frication burst of a labial stop is not shaped by a resonant cavity. If there is no coupling to the back cavity, the source flow spectrum $S(f)$ of a labial burst is radiated directly, with a transfer function of $T(f) = 1$.

Note that we use a different notation for front cavity resonances F_{fn} than for formants Fn. In this thesis, an indexed formant frequency Fn always refers to the nth resonance of the entire vocal tract, and front cavity resonances are differentiated by the subscript f. Since the front cavity is part of the vocal tract, the front cavity resonances are always a subset of the formants, $\{F_{f1}, F_{f2}, \ldots\} \subset \{F1, F2, \ldots\}$. The correspondence between the sets $\{F_{fn}\}$ and $\{Fn\}$ depends on the location of the consonantal constriction; this correspondence is discussed in more detail in section 2.3.

At the frequency of a vocal tract resonance, equation 2.17 can be approximated as

$$|T_{ap}(Fn)| \approx \frac{Fn}{Bn}\left(\prod_{j=1}^{n-1} \frac{Fj^2}{Fn^2 - Fj^2}\right) H(Fn) \tag{2.20}$$

In circuit theory, the first term on the right, Fn/Bn, is called the Q of the pole. The second term reflects the influence of lower-frequency formants, and the higher-pole correction $H(Fn)$ reflects the influence of higher-frequency formants.

The higher-pole correction can be large. If the vocal tract is uniform, circuit theory suggests that the higher pole correction is sufficiently large to make the amplitude of a formant peak $|T(Fn)|$ independent of the formant frequency Fn, and dependent only on the bandwidth Bn. Stevens (in preparation) has shown that the amplitudes of resonances in the transfer function of a uniform tube are approximately

$$|T_u(Fn)| \approx \frac{2S}{\pi Bn} \tag{2.21}$$

where S is the average formant spacing. If the vocal tract is not uniform, however, formant amplitudes depend significantly on the relative formant spacing. For non-uniform tubes, therefore, equation 2.21 may not be a good approximation, and equation 2.20 may be used to represent the influence of relative formant frequencies on amplitude.

2.2.2 Factors which influence formant bandwidth

Flanagan (1972) lists five sources of loss which contribute significantly to the formant bandwidths Bn, and therefore to the formant amplitudes: losses caused by viscosity, heat conduction, non-zero wall admittance, non-zero glottal admittance, and non-zero radiation

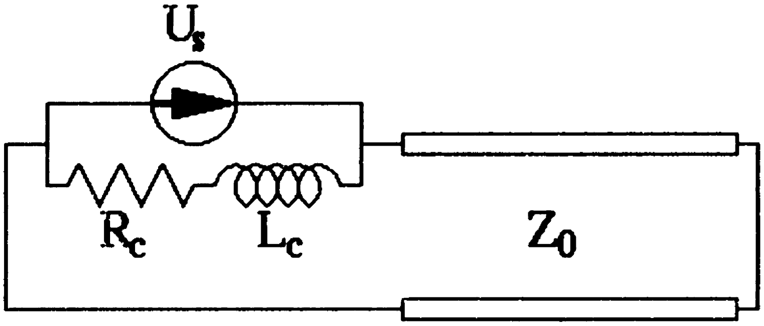


Figure 2-6: Circuit used for calculating the contribution of constriction losses to bandwidth. Resistance R_c and inductance L_c are shown in parallel with a constriction source flow U_s; vocal tract is shown as a transmission line with characteristic impedance Z_0.

impedance. Of these losses, only radiation losses and glottal losses (and constriction losses, which have the same form as glottal losses) will be important in this thesis.

The impedance of any constriction in the vocal tract, including a constriction at the glottis, can be represented by a resistance and inductance in parallel with a source flow U_s, as shown in figure 2-6. The resistance R_c and inductance L_c can be calculated, based on Bernoulli's equation and Newton' law, to be

$$R_c \approx \frac{k\rho U_0}{A_c^2}, \quad L_c = \frac{\rho l_c}{A_c} \tag{2.22}$$

where l_c and A_c are the length and area of the constriction, U_0 is the average flow through the constriction, and k is a constant that depends on the constriction shape, but is usually close to unity. If the vocal tract is modeled as a uniform tube, with a constant characteristic impedance Z_0, the contribution of constriction losses to the bandwidth of each formant is

$$Bn_c \approx \frac{R_c Z_0 c}{\pi l_t (R_c^2 + (2\pi Fn L_c)^2)} \tag{2.23}$$

where l_t is the length of the vocal tract.

Flanagan (1972) estimates that glottal losses typically add about 60Hz to the bandwidth of F1, and are less important for the higher formants. Klatt and Klatt (1990) suggest, however, that differences in glottal configuration from speaker to speaker may cause considerable variation in glottal losses; their measurements, as well as those of Hanson (1995), show a range of about 12dB in the implied bandwidth of F1. Phoneme-dependent differences in glottal configuration can also cause different amounts of loss. Equations 2.22 and 2.23 suggest that the contribution of glottal loss to the bandwidth of low-frequency formants is proportional to the square of the glottal area, so that if the glottal area doubles in size during aspiration, the bandwidth of F1 will quadruple.

The radiation impedance can also be represented by a series resistance and inductance, as shown in figure 2-7, but the radiation resistance turns out to be a function of frequency. If the mouth is modeled as a circular opening in a sphere, the radiation resistance is

$$R_r(f) = \frac{K(f)\pi\rho f^2}{c} \tag{2.24}$$

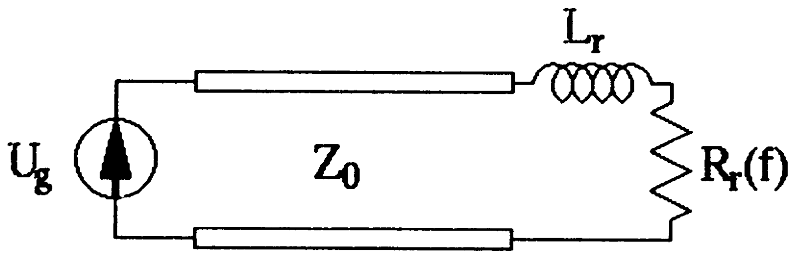


Figure 2-7: Circuit used for calculating the contribution of radiation losses to bandwidth. Radiation inductance L_r and resistance $R_r(f)$ are shown in series with the vocal tract transmission line.

where the correction term $K(f)$ is approximately 1.5 between the frequencies of 2000 and 6000 Hertz (Stevens, in preparation). If the vocal tract is modeled as a uniform tube, with a constant cross-sectional area A_t and characteristic impedance $Z_0 = \rho c/A_t$, the contribution of the radiation impedance to each formant bandwidth is

$$Bn_r \approx \frac{R_r c}{\pi Z_0 l_t} = \frac{K(f)(Fn)^2 A_t}{c l_t} \tag{2.25}$$

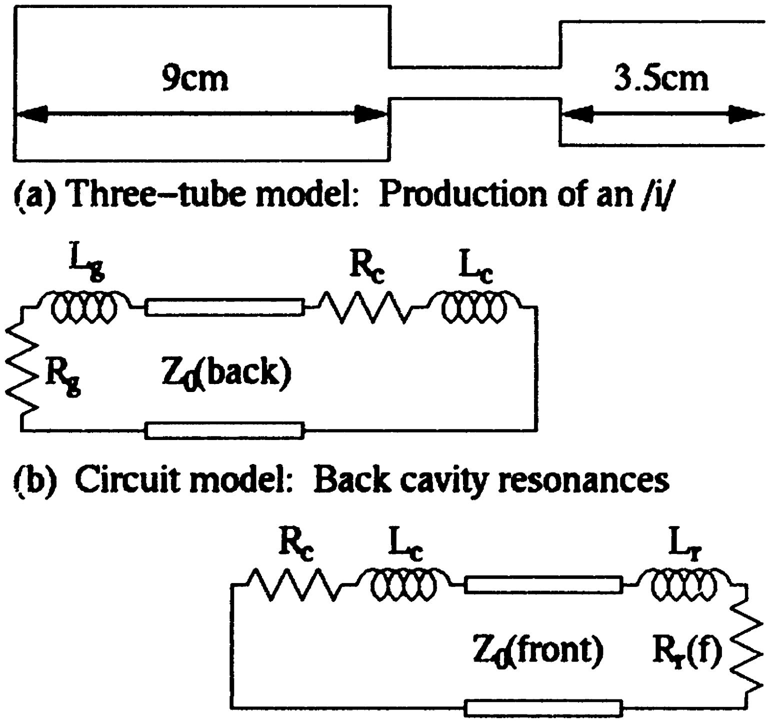
Given a 17cm vocal tract with an average area of 5cm^2, the contribution of radiation losses to bandwidth is roughly $(Fn/350)^2$ Hertz.

When the vocal tract is not uniform, the contribution of wall losses, glottal impedance, and radiation impedance to the bandwidth of each formant depends on details of the vocal tract shape. For example, consider the three-tube model of /i/ shown in figure 2-8a. The second resonance of this configuration is a half-wave resonance of the back cavity. The energy distribution of this resonance is strongly coupled to the glottal impedance, but almost completely decoupled from the radiation impedance; the bandwidth might be calculated using the circuit model shown in figure 2-8b, where the four-pole network includes models of wall losses in the pharynx region. The third formant is a quarter-wave resonance of the front cavity, and is therefore strongly coupled to the radiation impedance. The third formant is largely decoupled from the glottis, but instead of a glottal impedance, the bandwidth of the third formant is influenced by the constriction impedance, which can be modeled with the lumped element representation shown in figure 2-8c.

There is little recent work available on the relationship between vocal tract shape and formant bandwidths. The effect of vocal tract shape on formant bandwidth can be modeled quickly using available articulatory synthesizers (Maeda, 1982, Lin, 1990) , but these synthesizers have apparently never been used for a comprehensive study of bandwidth. Lin (1990) suggests that a comprehensive model of lower formant bandwidths would be difficult, given our current lack of knowledge about the distribution of wall losses inside the vocal tract. Higher formant bandwidths, on the other hand, depend primarily on the radiation impedance, for which we have fairly precise models (Flanagan, 1972), and should therefore be more susceptible to analysis by synthesis.

2.2.3 Pole-zero pairs

Often during the release of a consonant, the constriction at which acoustic sources are produced becomes large enough to allow coupling between the cavities in front of and



(a) Three–tube model: Production of an /i/

(b) Circuit model: Back cavity resonances

(c) Circuit model: Front cavity resonances

Figure 2-8: Three-tube model of the vocal tract, and circuit models of the back and front cavities, during production of an /i/. L_g and R_g are the glottal resistance and inductance, L_c and R_c are the constriction resistance and inductance, and L_r and $R_r(f)$ are the radiation resistance and inductance.

behind the source. When this happens, the spectrum is colored by local perturbations near the resonance frequencies of the back cavity.

Every resonance of a cavity upstream from the acoustic source contributes a complex pole pair and complex zero pair to the transfer function. A pole-zero pair can be modeled as a local multiplicative perturbation to the transfer function: at frequencies far from the pole-zero pair, the effect of the zero cancels the effect of the pole, and the total amplitude of the perturbation $T_{pz}(f)$ is a constant.

$$T_{pz}(f) = \frac{s_p s_p^*}{s_z s_z^*} \frac{(s - s_z)(s - s_z^*)}{(s - s_p)(s - s_p^*)} \tag{2.26}$$

When the constriction is completely closed, there is no coupling between the back cavity and the front cavity, and the pole frequencies $s_p = -\pi B_p + j2\pi F_p$ and zero frequencies $s_z = -\pi B_z + j2\pi F_z$ are exactly equal. When the constriction is slightly open, the back cavity pole and zero frequencies separate, and the pole, in particular, becomes more visible in the radiated spectrum.

The peak amplitude of the pole-zero perturbation can be approximated by assuming that the frequencies of the pole and zero are much larger than their bandwidths. Under this approximation, the peak amplitude of the perturbation is the Q of the pole, multiplied by a factor which depends on the separation of the pole and zero:

$$|T_{pz}(F_p)| \approx \frac{F_p}{B_p} \left| 1 - \left(\frac{F_p^2}{F_z^2} \right) \right| \tag{2.27}$$

The following sections discuss the frequencies and amplitudes of pole-zero pairs which may color the transfer function during aspiration and frication, respectively.

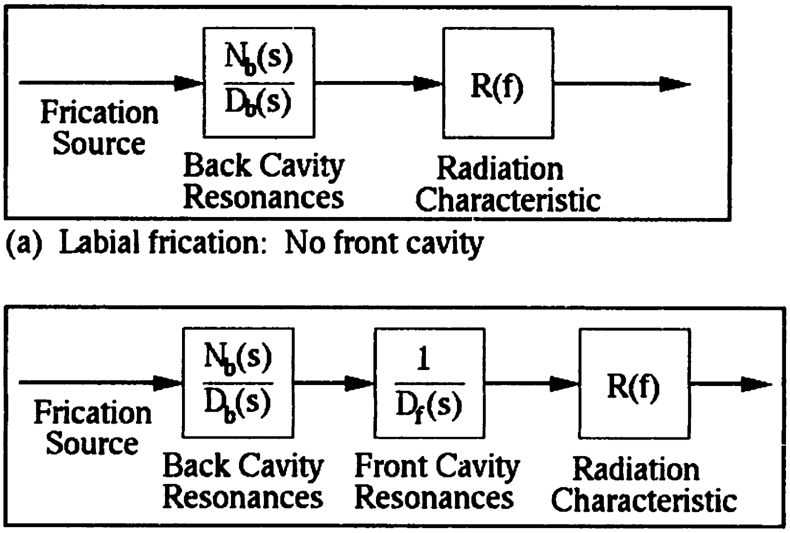
Subglottal resonances in aspiration

Ishizaka et al. (1976) measured the input impedance of the subglottal system on Japanese tracheotomized subjects. They found the first two resonances of the subglottal system to be at roughly 640 and 1400 Hertz, with Qs of roughly 10 and 18dB.

For a volume velocity source at the glottis, zeros in the transfer function occur at peaks of the subglottal impedance, which are usually very close to the measured resonance frequencies of 640 and 1400 Hertz. Poles of the transfer function occur at frequencies for which the sum of the subglottal, glottal, and supraglottal impedances is zero. In particular, it can be shown that the frequencies of the first two subglottal zeros are below the corresponding pole frequencies for any reasonable F1 and F2 (see Fant et al. 1972 for a discussion).

Given the pole and zero frequencies, the peak amplitudes of subglottal pole-zero perturbation functions $|T_{pz}(f)|$ can be calculated using the subglottal Qs found by Ishizaka et al., and the formula in equation 2.27. For example, if the first two pole-zero pairs in the vocal tract transfer function are at (800Hz,640Hz) and (1500Hz,1400Hz), as in the aspiration spectra of one of the subjects of Fant et al. (1972), the amplitudes of the transfer function perturbation $T_{pz}(f)$ at the frequencies of the two poles are roughly $|T_{pz}(800)|$ =5dB and $|T_{pz}(1500)|$ =1.5dB.

Under normal circumstances, an LPC formant tracker ignores subglottal resonances, because the poles do not contribute to the global shape of the spectrum. Section 2.2.5 considers circumstances under which subglottal resonances may interfere with LPC measurement of the vocal tract formants.



(a) Labial frication: No front cavity

(b) Alveolar/Velar frication: Front and back cavities

Figure 2-9: Source-filter models of frication at the lips, and of frication at a tongue constriction.

Back cavity resonances in frication

We have previously described the transfer function of a frication burst, from equivalent source flow $U_s(f)$ to mouth flow $U_m(f)$, as being an all-pole spectrum, with peaks at the front cavity resonant frequencies F_{fn} (section 2.2.1). Often, however, the constriction is large enough to allow coupling to the back cavity, and resonances of the back cavity cause pole-zero perturbations in the transfer function.

In a model which allows coupling to the back cavity, the transfer function of a labial stop is entirely composed of pole-zero pairs, as shown in figure 2-9a. There is no front cavity, but there are back cavity resonances at every formant frequency.

Frication produced at a tongue blade or tongue body constriction is filtered by poles and zeros associated with the back cavity, and also by poles associated with the front cavity (figure 2-9b). Zeros occur at peaks of the back cavity impedance, that is, at resonant frequencies of the back cavity. Poles occur at frequencies for which the sum of the back cavity, constriction, and front cavity impedances is zero, that is, at the resonant frequencies of the vocal tract—the formant frequencies.

The amplitude of the spectral peaks corresponding to each back cavity resonance depend on $Q = F_{pn}/B_{pn}$, and the ratio F_{pn}/F_{zn} of the frequencies of the spectral pole and zero. Empirically, labial and velar stops produced with a palatal constriction (in syllables like "keel" and "pyew") often have well-separated pole-zero pairs near the frequency of the constriction resonance, which is typically F3 or F4. Alveolar stops which are released quickly may have strong pole-zero pairs at the frequencies of F2 and F3.

2.2.4 Nasalization

In order to pronounce a nasal consonant with no loss of sonorant voicing, the velopharyngeal port must be opened prior to oral closure, and kept open until after the oral constriction is released. During closure, or when the oral constriction is smaller than the opening of the velopharyngeal port, the transfer function is dominated by resonances of the nasal-pharyngeal system (Fujimura, 1962). When the velopharyngeal port is smaller than the oral constriction, but still open, the transfer function is dominated by the oral formants, but there are still pole-zero pairs near the frequencies of the nasal-pharyngeal resonances. Chen (1991) has documented the presence of pole-zero pairs at about 300 Hertz and 1000 Hertz in the speech of hearing impaired subjects, corresponding to the first two nasal resonances reported by Fujimura.

The spectrum of a vowel immediately after release of a nasal consonant can be approximately modeled as an all-pole transfer function $T_{ap}(f)$, modified by pole-zero perturbations $T_{pz}(f)$ at approximately 300 and 1000 Hertz. The pole-zero perturbations continue to color the spectrum until the velopharyngeal port closes. Typically, the velopharyngeal port closes within 10-20ms after release of a nasal consonant, but there is considerable variability, depending primarily on segmental and prosodic context.

After release of a nasal consonant, pole-zero perturbations may interfere with the measurement of formant frequencies. Specifically, the large pole-zero perturbations at about 300Hz and 1000Hz are likely to interfere with measurement of F1, and possibly of F2 as well. This interference will only affect formant measurements while the velopharyngeal port is open, and therefore typically only for the first 10-20ms following release.

2.2.5 Measurement issues: LPC

Linear predictive coding, or LPC (Atal and Hanauer, 1971) is the commonly used name for a group of algorithms which efficiently compute the poles of an all-pole spectrum. If any spectrum can be modeled as the product of a flat source spectrum and an all-pole transfer function, $X(f) \propto T(f)$, LPC can be used to estimate the resonances of the transfer function $T(f)$.

LPC assigns pole frequencies in order to minimize an error term. This section analyzes the error term as the product of a local term representing the fit to an individual peak, and a global term representing the global spectral fit. Based on this analysis, an original qualitative analysis is given of the conditions under which large formant frequency errors are expected to occur.

Local and global spectral error terms

Rabiner and Schafer (1978) show that the LPC algorithm finds a unity-gain, all-pole estimate $\widehat{T(f)}$ of the spectrum $X(f)$ which minimizes the error term

$$E \propto \int_0^{f_s} \frac{|X(f)|^2}{|\widehat{T(f)}|^2} df \tag{2.28}$$

where f_s is the sampling frequency, and

$$\widehat{T(f)} = \prod_{i=1}^{m} \frac{\widehat{s}_i \widehat{s}_i^{\,*}}{(s - \widehat{s}_i)(s - \widehat{s}_i^{\,*})} \tag{2.29}$$

The error term in equation 2.28 is a global spectral error, but since the radiated spectrum is in the numerator, a given fractional error near a peak of $X(f)$ is weighted more heavily than the same fractional error near a valley. The result is that each pole estimated by LPC is a compromise between a global spectral fit, and a local fit to a single spectral peak.

If the order of the model is chosen correctly, the estimated formant frequencies $\widehat{Fn}$ are usually close to the true formants Fn. In evaluating the possibility of a formant frequency error, it is instructive to evaluate the LPC error term at the formant frequencies. If $T(f)$ is all-pole with the form given in equation 2.17, the integrand in equation 2.28 near a formant frequency is approximately

$$\frac{|T(Fn)|^2}{|T(\widehat{Fn})|^2} \approx \left(\frac{\sqrt{4(Fn-\widehat{Fn})^2+\widehat{Bn}^2}}{Bn}\right)^2 \left(\prod_{j=1}^{n-1}\frac{(Fj^2)(Fn^2-\widehat{Fj}^2)}{(\widehat{Fj}^2)(Fn^2-Fj^2)}\right)^2 \left(\frac{H(Fn)}{H(\widehat{Fn})}\right)^2 \qquad (2.30)$$

Here, the error integrand has been separated into a local error term, expressing the dependence of error on bandwidth, and two global error terms, expressing the dependence of error on the relative formant frequencies. The first term on the right hand side of the equation shows the local constraint: the estimated formant frequency is loosely constrained within about half a bandwidth of the true formant, but if the difference $|Fn - \widehat{Fn}|$ is larger than about half a bandwidth, the error increases sharply. The second and third terms show the global constraint: the relative formant positions of the model $\widehat{T(f)}$ must be similar to those of the observed spectrum.

Subglottal resonances

Equation 2.30 shows that LPC usually ignores a pole-zero perturbation in the spectrum, even if the amplitude of the pole is large, because if LPC assigns too many complex pole pairs to any given frequency band, the relative spacing of the formant frequencies will be incorrect. According to this reasoning, LPC only assigns a complex pole pair to a spectral perturbation if it also fails to assign a pole pair to a nearby formant. Since the LPC error metric weights spectral errors by the amplitude of the DFT spectrum, a spectral perturbation usually does not "steal" a complex pole pair in this manner unless the perturbation amplitude is greater than the amplitude of the neighboring formant peak.

We have seen in section 2.2.3 that under normal circumstances, the amplitude of a pole-zero spectral perturbation is usually considerably less than the amplitudes of nearby formants. For a perturbation to have a higher amplitude than a nearby formant, the formant must be unusually weak. There are several reasons why this might happen; this section discusses three.

First, during aspiration, a formant may be temporarily wiped out by a spectral null caused by random variation in the source spectrum. According to the calculations in section 2.1.4, roughly one out of every 100 independent samples of the time-averaged power spectrum is affected by a random spectral null of -11dB or more. If the spectrum is calculated using a 6ms Hanning window, the bandwidth of the spectral null is $2/0.006 \approx 330$ Hertz, which is sufficiently wide to temporarily wipe out a formant.

Second, the bandwidth of the formant may become so large that there is no resonant peak at the frequency of the formant. This is most often a problem with the F1 peak during aspiration, because of the large glottal area. If, for example, the average glottal area doubles during aspiration, then the glottal inductance L_g and resistance R_g in equation 2.23 decrease by factors of 2 and 4, respectively. Decreasing R_g by a factor of 4 increases the first formant bandwidth $B1$ by a factor of 4, to more than 200 Hertz, which is sufficiently large to nearly wipe out the F1 resonant peak in the spectrum. The enlarged glottal area in aspiration also provides increased separation of the first subglottal pole and zero, with the result that the first subglottal pole is often more prominent than F1 in the aspiration following release of an unvoiced stop.

Third, two formants which are close together may merge into a single spectral peak. The resonance curve of an excited formant always contributes to the spectrum, but a nearby formant with a higher amplitude (and narrower bandwidth) often tilts the resonance curve so severely that there is no convex peak at the frequency of the lower-amplitude formant. For convenience, this phenomenon can be referred to as "formant merger," where the lower-amplitude formant is said to merge with the higher-amplitude formant. If there is a subglottal resonance near the merged formants, or if the global spectral shape is ambiguous, LPC occasionally assigns a single complex pole pair to the merged formants. Hillenbrand, Getty, Clark and Wheeler (1995) found that about 3% of their vowel nuclei contained an F2-F3 merger which could not be resolved by interactively changing the LPC analysis order, while about 1% of tokens contained an unresolvable F1-F2 merger.

A subglottal resonance which is ignored by LPC analysis may still cause problems for formant tracking. If a moving formant frequency crosses the frequency of a pole-zero perturbation, linear system theory predicts that the frequency of the formant skips discontinuously across the frequency of the subglottal resonance (Hanson and Stevens, 1995). The size of the discontinuity depends on the relative amplitudes of the subglottal and supraglottal impedance, and is typically between about 200 and 300 Hertz.

2.2.6 Measurement issues: frication spectrum

The burst spectrum often contains spectral perturbations at the frequencies of back cavity and even subglottal resonances, as well as peaks corresponding to the front cavity resonances. This section discusses and compares factors influencing the amplitudes of front and back cavity resonances in a frication spectrum. As part of this discussion, original quantitative limits on the amplitudes of front and back cavity resonance amplitudes are derived.

The discussion below focuses on amplitudes of the transfer function between an equivalent source flow $U_s(f)$ and the mouth flow $U_m(f)$. To compute actual radiated spectral amplitudes, the transfer function amplitudes discussed below must be multiplied by the amplitudes of the equivalent flow source $U_s(f)$ and the radiation characteristic $R(f)$. As discussed in section 2.1.2, there is usually a downward tilt in the spectrum $U_s(f)R(f)$ at high frequencies, but the tilt depends significantly on the relative amplitudes of the monopole and dipole frication sources, and therefore on the shape and position of the constriction.

At very low frequencies, the frication spectrum is dominated by the monopole source, and is not at all well modeled by a simple spectral tilt. Figure 2-4a indicates that the monopole frication spectrum has a concentration of energy below about 700Hz (although this is difficult to see, because the figure only shows frequencies above 500Hz). Empirically, radiated frication spectra often contain one or more large peaks at low frequencies, typically below about 700Hz.

Front cavity resonances

Transient and frication sources at the release of a stop excite the resonances of the front cavity, with local perturbations caused by resonances of the back cavity and constriction. If the front cavity is modeled as a uniform tube, the resonance frequencies are as reviewed in section 2.2.1:

$$F_{fn} = \frac{(2n-1)c}{4l_f}, \quad |T_{ap}(Fn)| \approx \frac{2S_f}{\pi B_{fn}} \tag{2.31}$$

where $S_f = c/2l_f$ is the average front cavity resonance spacing, and $T_{ap}(Fn)$ is the all-pole component of the transfer function.

The bandwidths of high-frequency front cavity resonances depend largely on radiation losses, while the bandwidths of low-frequency resonances may be controlled by losses at the constriction. Since radiation losses are proportional to F_{fn}^2, equation 2.31 predicts that the transfer function amplitude of a high-frequency front cavity resonance should fall as $1/F_{fn}^2$. If we ignore the constriction inductance, the contribution of the constriction to bandwidth can be approximated as a fixed constant,

$$B_c \approx \frac{Z_0 c}{\pi l_f R_c} \tag{2.32}$$

where Z_0 is the characteristic impedance, and R_c is the constriction resistance. With this approximation, the amplitude of resonance peaks in the transfer function can be written:

$$|T_{ap}(F_{fn})| \approx \frac{c^2}{\pi K(f) A_f (f_c^2 + F_{fn}^2)}, \quad f_c \approx \sqrt{\frac{c^2 Z_0}{\pi K(f) A_f R_c}} \tag{2.33}$$

The cutoff frequency f_c varies considerably depending on the area of the constriction and the area of the front cavity, but is typically in the mid-frequency range. If, for example, $R_c \approx 100$ acoustic ohms, and $A_f \approx 5cm^2$, f_c is approximately 2000Hz.

If the front cavity is modeled as a uniform tube, equation 2.31 specifies that all of the front cavity resonances are odd multiples of the first resonance frequency F_{f1}. In this case, equation 2.33 can be used to compute the relative amplitudes of transfer function resonance peaks in different frequency bands, regardless of the exact resonance frequencies. If F_{f1} is much larger than f_c, for example, equation 2.33 predicts that $|T_{ap}(F_{f2})|$, the transfer function amplitude of the second front cavity resonance, is $20\log(9)$ =19dB below $|T(F_{f1})|$. If the cutoff frequency f_c is somewhere between F_{f1} and F_{f2}, as might be true at the release of a velar stop, the difference between their transfer function amplitudes will be somewhere between 0 and 19dB. The differences in amplitude of the radiated spectral peaks will typically be somewhat less, because the product $R(f)U_s(f)$ usually has a slightly positive tilt during frication.

Back cavity resonances

Resonances of the subglottal system, the back cavity, and the oral constriction often contribute multiplicative pole-zero perturbations to the transfer function. The amplitude of the transfer function at the frequency F_p of a back cavity resonance is approximately the amplitude of the pole-zero perturbation $T_{pz}(f)$, as given in equation 2.27:

$$|T_{pz}(F_p)| \approx \frac{F_p}{B_p}\left(\left(\frac{F_p}{F_z}\right)^2 - 1\right) \tag{2.34}$$

The amplitude of the perturbation can be estimated if we recognize that, as discussed in section 2.2.3, the poles are at formant frequencies of the entire vocal tract, while the zeros are at resonant frequencies of the back cavity. For poles and zeros below the frequency of the first front cavity resonance, it can be shown that the pole and zero frequencies are interleaved, $F_{z1} < F_{p1} < F_{z2} < F_{p2} < \ldots$, so that the spacing between a pole and zero will never be larger than half the spacing between adjacent formants. If the average spacing of vocal tract formants is S, we can use the inequality $F_p \leq F_z + S/2$ in equation 2.34 to get

$$|T_{pz}(F_p)| \leq \frac{S(S+4F_z)}{4B_pF_z} \approx \frac{S}{B_p} \tag{2.35}$$

Section 2.2.1 mentions that the amplitude of a formant peak in a vowel transfer function, assuming a uniform vocal tract, is approximately $2S/\pi B$. Essentially, then, equation 2.35 shows that the transfer function amplitude of a back cavity resonance should be less than or equal to the transfer function amplitude of the corresponding formant in the following vowel.

The relative amplitudes of back cavity and front cavity resonances can be computed by combining equations 2.35 and 2.31:

$$\frac{|T_{pz}(F_p)|}{|T_{ap}(F_{fn})|} \leq \frac{(S/B_p)}{(2S_f/\pi B_{fn})} \tag{2.36}$$

Since the vocal tract formant spacing S is much less than the front cavity spacing S_f for most consonantal configurations, equation 2.36 demonstrates that the amplitude of a back cavity resonance will be much less than the amplitude of any front cavity resonance with a similar bandwidth.

The bandwidth of a low-frequency front cavity resonance is controlled by constriction losses (see equation 2.33), and is therefore similar in size to the bandwidths of back cavity resonances. The bandwidth of a high-frequency front cavity resonance, however, is controlled by radiation losses, and, at very high frequencies, may be significantly larger than the bandwidths of some back cavity resonances. Thus we find that the transfer function amplitude of a front cavity resonance is larger than the transfer function amplitudes of any back cavity resonances, unless the front cavity resonance is at very high frequency. The words "very high frequency" will be made slightly more quantitative in the empirical study of section 3.3.3.

2.3 Evolution of Vocal Tract Resonances at Release of a Consonant

This section describes, for each class of consonants, the theoretical basis for predicted correlations between consonant place and measurements of the vocal tract resonances. This section describes the correlations we expect to see in the measurements later in this thesis, but it is not strictly necessary as background reading for the thesis.

2.3.1 First formant trajectory

For all three places of articulation, the first vocal tract resonance at release is a Helmholtz resonance, which can be calculated by modeling the back cavity as an acoustic capacitor, and the front cavity as an inductor in parallel with the inductance of the vocal tract walls (figure 2-10). Given the length and area of the constriction l_c and A_c, and the volume of the back cavity V_b, the first formant frequency can be approximated as

$$F_1 \approx \frac{1}{2\pi}\sqrt{\frac{L_w + L_c}{L_wL_cC_b}} = \frac{1}{2\pi}\sqrt{\frac{1}{L_wC_b} + \frac{A_cc^2}{l_cV_b}} \tag{2.37}$$

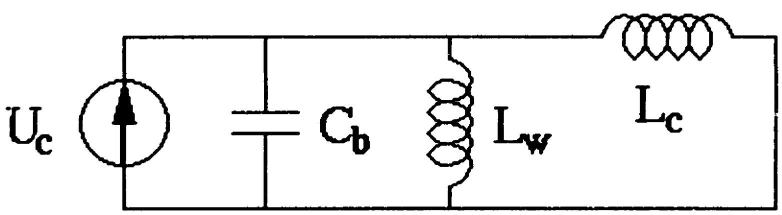
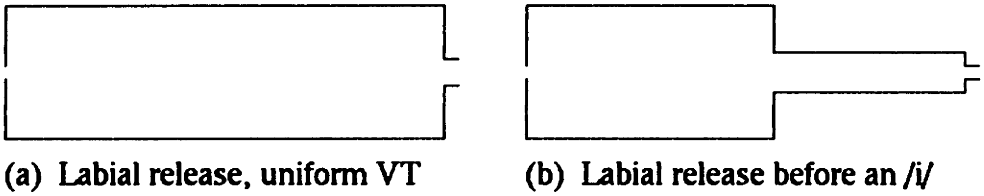


Figure 2-10: Low-frequency lossless circuit model of the vocal tract after releasė of a stop, used for calculating the first formant frequency. The back cavity is treated as a lumped capacitor C_b, which resonates with the parallel inductances of the yielding wall, L_w, and stop constriction, L_c.


(a) Labial release, uniform VT

(b) Labial release before an /i/

Figure 2-11: Simplified tube models of the vocal tract immediately after release of a labial stop, with different shapes of the back cavity.

where, in the second step, the lumped element values $L_c = \rho l_c/A_c$ and $C_b = \rho c^2/V_b$ have been used. The resonant frequency of the closed vocal tract, $1/2\pi\sqrt{L_wC_b}$, has been measured at about 220 Hertz.

Equation 2.37 predicts that, after F1 begins to rise away from the closed-tract resonance, the frequency of F1 rises in proportion to the square root of the constriction area. The constant of proportionality depends on the shape of the constriction. Labial and alveolar consonants, for example, have short constrictions (typically 1-2cm), so equation 2.37 predicts that F1 should rise quickly at release of a labial or alveolar consonant. The constriction of a velar consonant is typically 3-5 times as long as that of a labial or alveolar, so equation 2.37 predicts that F1 should rise about half as fast. Empirically, most of the F1 transition usually occurs within about 20 ms following the release of a labial or alveolar stop, and within about 50 ms following the release of a velar stop.

2.3.2 Labial releases

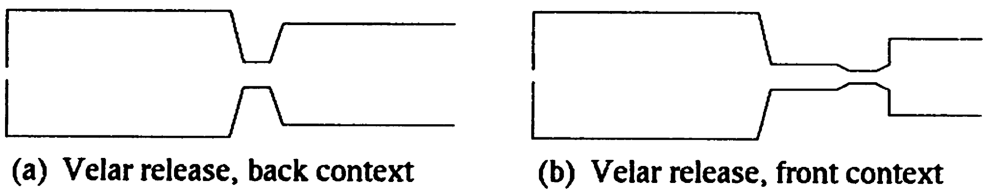
The resonances of the vocal tract immediately after release of a labial consonant can be estimated using a model similar to that shown in figure 2-11a. The back cavity, consisting of the entire vocal tract, is nearly closed at both ends; there is no front cavity.

If the back cavity is modeled as a uniform tube, the formants F2,F3,... immediately after release take the values

$$Fn = \frac{nc}{2l} \tag{2.38}$$

If the vocal tract length l is about 17cm, the formants at onset are roughly F2=1040 Hertz, F3=2080 Hertz, etc. As the constriction area increases, these formant frequencies rise toward the vowel target formants.

During labial closure, the tongue often moves toward the configuration of the following


(a) Velar release, back context

(b) Velar release, front context

Figure 2-12: Simplified tube models of the vocal tract immediately after release of a velar stop, with backed and fronted tongue positions.

vowel, so that the back cavity may be significantly non-uniform at release. Changes in the back cavity shape can affect both the onset frequency and rate of change of formants at release of a labial stop.

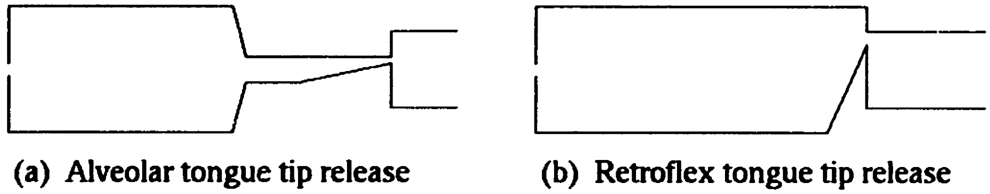
Manuel and Stevens (1995), for example, have modeled a labial stop followed by a high front vowel using a model similar to that shown in figure 2-11b. In this model, the palatal constriction resonance starts at $c/4x$, where x is the length of the constriction. By the time the lip area is the same as the constriction area (typically about 10ms after release), the palatal constriction resonance doubles in frequency to $c/2x$, while the resonance of the back cavity remains nearly unchanged. If the constriction is 7cm in length and the back cavity is 10cm, for example, the constriction resonance rises from about 1300 to about 2500 Hertz, while the back cavity resonance is constant at about 1800 Hertz.

2.3.3 Velar releases

At release of a velar consonant, the vocal tract resonances can be divided into resonances of the front cavity, back cavity, and constriction. The actual lengths of the front cavity and constriction depend significantly on the following vowel. This section describes velar stops using the traditional distinction between backed and fronted constrictions (e.g. Halle, Hughes, and Radley, 1957), because this traditional distinction is a useful predictor of the cavity affiliations of F2 and F3 (see below). Lehiste and Peterson (1961), however, argue that the location of an English velar closure varies continuously depending on the following vowel, and that the front and back allophones described below should be considered as points on a continuum, rather than well-separated phonetic categories.

If the consonant is followed by a typical back vowel (e.g. /aa/ or /ah/), the stop closure can be modeled using a tube model similar to the one shown in figure 2-12a. In this model, the front cavity is roughly 6-8cm long, and the first two front cavity resonances are at 1100-1500 and 3300-4500 Hertz. The transfer function amplitude at the frequency of the first resonance is controlled by constriction losses, while the amplitude of the second resonance is controlled by radiation losses, as discussed in section 2.2.6. The relative amplitudes will vary considerably depending on the areas of the front cavity and of the constriction. As a representative example, if the front cavity area is 5cm^2, and the constriction resistance is about 80 acoustic ohms, then the transfer function amplitudes of the first two resonances are roughly 20dB and 6-10dB, depending on frequency. The peaks in the radiated sound spectrum are at these amplitudes, scaled by the radiation characteristic, and by a linear combination of the monopole and dipole frication sources discussed in section 2.1.2.

In back vowel context, F2 usually starts at the frequency of the first front cavity resonance, while F3 starts at the half-wave resonance of the back cavity. If the constriction is about 1cm long, for example, and for a 17cm vocal tract, the front cavities given above


(a) Alveolar tongue tip release

(b) Retroflex tongue tip release

Figure 2-13: Simplified tube models of the vocal tract immediately after release of alveolar and retroflex stops.

would correspond to back cavities of 8-10cm in length, with F3 onset frequencies of roughly 1800-2200 Hertz.

A model of a velar release before a front vowel is shown in figure 2-12b. Before a front vowel, velar stops in English tend to become palatalized, with a long palatal constriction (typically 4-5cm) and a short front cavity (typically 3-4cm). The first front cavity resonance is typically 2000-3000 Hertz, and has a transfer function amplitude of 17-23dB. Because of the high impedance of the constriction, the separation F_p/F_z of the zero and pole associated with the first constriction resonance is often large, and the pole (typically at about 4000 Hertz) is often prominent in the frication spectrum.

In front vowel context, F3 usually starts at the frequency of the first front cavity resonance, and F2 starts at the half-wave resonance of the back cavity. Empirically, the first back cavity resonance is in the same range (roughly 1800-2200 Hertz) regardless of whether the velar is backed (in which case the resonance becomes F3) or fronted (in which case it becomes F2).

2.3.4 Alveolar, retroflex, and lateral releases

Production of an alveolar constriction (figure 2-13a) constrains the position of the tongue more than does production of a velar and labial constriction. The length of the front cavity, between the tongue tip and the lips, is usually 1-2cm in length (depending on the degree of lip rounding); if the radiation inductance is taken into consideration, the effective front cavity length is roughly 1.5-2.5cm. The tongue body is always fronted to support the tongue tip. At the release of an alveolar, F2 is approximately a half-wave resonance of the back cavity. Constraints on the tongue body position (Manuel and Stevens, 1995) cause the onset frequency of F2 after an alveolar stop to be considerably less variable than the F2 onsets of velar and labial stops: this frequency is typically 1500-1900 Hertz for male speakers, and 1900-2200 Hertz for females (Sussman et al., 1991).

The onset frequency of F3 is approximately a full-wave resonance of the back cavity. Basic acoustic theory predicts that the frequency of a full-wave back cavity resonance should be twice that of the half-wave resonance. Empirically, the onset frequency of F3 after alveolars is higher than it is at labial and velar releases, but is usually less than twice the frequency of F2.

Assuming a front cavity of 1.75-3cm in length, the first front cavity resonance in the frication transfer function is between 3000 and 5000 Hertz (usually the onset frequency of F4 or F5). If the cross-sectional area is about 3cm^2, the transfer function amplitude at the resonant frequency is between 12 and 21dB. When an alveolar stop is released quickly, transient ringing or back cavity coupling often introduces large perturbations in the burst spectrum. Stevens and Blumstein (1979) found large back cavity resonances at the frequency of the first subglottal resonance (700-1000 Hertz), and at the onset frequency of F2.

The shape and position of an alveolar constriction is rarely influenced by the features of a neighboring vowel, but a neighboring phoneme which requires a specific tongue *blade* configuration often dramatically changes the shape of an alveolar consonant. When an alveolar consonant is followed (or sometimes preceded) by a retroflex glide or vowel (/r/ or /er/), the tongue blade closure during the consonant moves back behind the alveolar ridge, effectively adopting the retroflex place of articulation. When an alveolar consonant is followed by a lateral phoneme (/l/ or /el/), the closure is often released on one side of the tongue, rather than at the tip, effectively adopting the lateral place of articulation.

A retroflex stop (figure 2-13b) has a much longer front cavity than an alveolar stops. In the neighboring retroflex vowel or liquid, F3 is usually a front cavity resonance, often with a frequency below 2000 Hertz (Peterson and Barney, 1952), corresponding to a front cavity length of about 4.5cm. At the release of a stop or nasal which has assimilated retroflex articulation, the front cavity may be slightly shorter than 4.5cm; empirically, the front cavity resonance of a retroflex stop tends to be between 2000 and 3000 Hertz. The onset value of F3 tends to be associated with the front cavity resonance, while the onset frequency of F2 is associated with the back cavity.

Lateral vowels and liquids are characterized by a low F2, a possible pole-zero pair at the usual frequency of F3, and a cluster of resonances near the usual frequency of F4 (Stevens, in preparation). If an alveolar stop is released directly into a lateral configuration, the formant frequencies at onset usually match the formant frequencies of a typical lateral, and may have little relationship to the typical onset formants of an alveolar.

2.4 Summary and Discussion

This chapter has developed production models to explain, first, the correlation between resonant frequency measurements and the place of articulation of a consonant, and second, the types of source and filter variability which make resonant frequencies difficult to measure. These discussions have been intended to serve two purposes. First, discussions in this chapter provide specific background for the design of acoustic correlate measurements in chapter 3. Second, the models of variability in this chapter are intended as general examples of the power of speech production modeling to explain acoustic variability.

2.4.1 Acoustic correlates of place

In the design of algorithms to measure formant and front cavity resonance frequencies, chapter 3 will refer to several of the speech production models developed or reviewed in this chapter. This section reviews briefly some of the important results from this chapter which will be used again in chapter 3.

Burst spectrum

The burst spectrum of a stop is primarily shaped by the resonances of the front cavity, if there is one. The burst for a labial stop, with no amplification from front cavity resonances, is usually lower in amplitude than that for an alveolar or velar stop. The front cavity resonance frequency of an alveolar or velar stop, if measured correctly, almost always determines the place of the stop.

Errors in measurement of the front cavity resonance may be caused by back cavity resonances, which appear as pole-zero pairs in the transfer function of a fricative burst. In general, the burst spectra of quickly released stops, usually labials and alveolars, contain more evidence of back cavity resonances than the spectra of more slowly released stops, usually velars, although there is considerable variability depending on phonetic context.

The transfer function amplitude of a front cavity resonance is usually higher than that of back cavity resonances in the same spectrum. Because of radiation losses, however, the bandwidth of a high-frequency front cavity resonance can be quite wide, with the result that very high frequency front cavity resonances will occasionally be lower in amplitude than back cavity resonances in the same spectrum.

The spectrum of a fricative burst is also shaped by variation in the source spectrum. The relative amplitudes of high-frequency and low-frequency peaks vary considerably depending on the relative amplitudes of the monopole and dipole turbulent sources. Often, there will also be a large peak below about 700 Hertz, corresponding to low-frequency energy in the monopole turbulent source.

Formant motion

Knowledge of formant motion into the following vowel can provide information about consonant place.

During aspiration, the first subglottal resonance is often mislabeled as F1, so that measurements of F1 in aspiration are often meaningless. F1 may also be obscured at the release of a nasal consonant, because of pole-zero pairs corresponding to resonances of the nasal-pharyngeal system.

Higher formants are also difficult to track during aspiration. Variations in the source spectrum may occasionally zero out a formant with a randomly generated spectral null. Subglottal resonances may contribute pole-zero pairs to the spectrum, which may, under certain conditions, be mistakenly identified as formants. Finally, even if a formant is tracked correctly, the spectral peak corresponding to a formant which crosses the frequency of a subglottal pole will occasionally skip discontinuously across the pole, with an instantaneous discontinuity of 200-300 Hertz.

2.4.2 Production models of variability

This chapter has developed or reviewed production models of variability in the turbulent source spectrum, of bandwidth variation and subglottal resonances in aspiration, and of front and back cavity resonances in frication.

The introduction to this chapter proposed the hypothesis that speech production models can help to explain the aggregate statistical distribution of acoustic correlate measurements. Based on the examples in this chapter, it is now possible to discuss limitations and implications of the hypothesis in more detail.

Production models can be used with different degrees of success to model different kinds of variability. Random or chaotic variation in the source spectrum, for example, is caused by a nonlinear physical process which varies little from speaker to speaker. In this case, since variability is generated by a known physical process, it is possible to derive an explicit probability density based entirely on physical models of turbulent flow (the χ^2 model developed in section 2.1.4).

Variability in the transfer function amplitude of back cavity resonance peaks, on the other hand, was shown to depend on the separation of the associated pole and zero frequencies, which, in turn, depends on the speed of stop consonant release. It is therefore only possible to derive a theoretical distribution of back cavity spectral amplitudes if we already happen to know the distribution of stop consonant release rates. The distribution of release rates, however, depends at least partly on decisions and physical characteristics of the speaker.

With the exception of turbulent source variation, all of the acoustic correlate variability discussed in this chapter depends on anatomical and stylistic differences between speakers. The bandwidth of F1, for example, depends on the area of the glottis, which is governed by both anatomy and speaking style. The bandwidth of front cavity resonances depends on the area of the constriction and the area of the lip opening, both of which are under the stylistic control of the speaker.

When the distribution of an acoustic correlate is under the control of the speaker, the predictive power of theoretical models is limited. A detailed probability distribution can only be obtained from empirical measurements of either the acoustic correlate (e.g. back cavity spectral peaks) or the articulatory correlate (e.g. stop consonant release rate). In most cases, the acoustic correlate is easier to measure than the articulatory correlate.

Production models can help to explain the variability of speaker-controlled parameters in two ways. First, a production model can predict physical limits on the range of variability. Sometimes these limits are extremely weak, but chapter 3 will demonstrate that even an extremely weak bound on the expected variability can be helpful in the design of acoustic correlate measurements.

Second, a production model can be used to evaluate measured acoustic correlate distributions, by defining a relationship between acoustic and articulatory parameters. Since acoustic correlates are usually easier to measure than articulatory correlates, it is probably easier to predict the distribution of the articulatory measurement from the distribution of the acoustic correlate, rather than the other way around. If predictions of articulatory variation based on one acoustic correlate are confirmed by predictions from other acoustic correlates, or by direct articulatory measurements, this confirmation then helps to develop our knowledge of speech production.

The development of models of variability in this chapter, therefore, leads to the suggestion of a particular methodology for learning more about speech production, in which measured distributions of acoustic parameters are combined with speech production knowledge to predict equivalent articulatory distributions. If the acoustic measurements can be automated, the proposed method promises rapid collection and confirmation of detailed descriptions of speech production variability.

The next chapter describes the development and evaluation of rule-based algorithms for the measurement of certain acoustic parameters related to the place of a consonant release. In the original plan for this thesis, these algorithms were designed for the purpose of learning about speech variability, using the methodology described above. Unfortunately, it turns out that the measurement algorithms developed in chapter 3 are prone to measurement error. Rather than predict articulatory distributions on the basis of erroneous acoustic measurements, chapter 3 turns instead to the more conservative task of modeling the distribution of measurement error, using both statistical and speech production knowledge to help build our understanding of the sources and characteristics of acoustic measurement error.

Chapter 3

Rule-Based Measurement Algorithms

This chapter describes a knowledge-based approach to minimizing and characterizing error in the measurement of formants and front cavity resonances near the release of a stop consonant. The output of a commercial formant tracker, and time-averaged power spectra of the burst, are modified and searched, by rule, to imitate the measurements of a human judge on a training set; the resulting algorithms are referred to in this thesis as rule-based algorithms. The measurement error of the algorithm, as compared to a human transcriber, is then modeled in the same way that differences between two human judges are measured: errors on a test set are measured, and, depending on the number of tokens in the test set, confidence limits are computed for the error mean, error variance, and probability of large errors.

The explicit goal of this chapter is to design formant and burst spectral measurements which imitate, as closely as possible, the measurements made by a human judge. The implicit goal is to create algorithms which can be used in every application for which manual measurements have previously been used, specifically, in speech sound classification (introduced in section 1.1, and covered in more detail in chapter 5) and for inference of the distribution of articulatory variables (introduced in section 2.4.2).

Phonetic classification and articulatory inference require detailed knowledge about any possible errors in an acoustic measurement. This chapter will develop four types of error model. First, the mean and standard deviation of a simple additive Gaussian error model will be calculated. Errors in the measurement of frequency parameters will be shown to contain outliers not well modeled by a Gaussian; these outliers will be modeled using mixture Gaussian models, and using a non-parametric analysis of the phonetic contexts in which they occur. Finally, errors in the measurement of amplitude parameters, and of F1 at stop release, will be shown to be correlated with the correct value of the measurement, and therefore to require a heteroskedastic error model.

3.1 Signal Representation

In order to draw as much as possible on the prior experience of the judges, the measurements in this chapter are formulated in terms of signal representations typically used in interactive phonetic analysis.

3.1.1 LPC-based formant frequencies

In many experiments in this thesis, estimated formant frequencies during aspiration and voicing are based on the roots of an LPC polynomial, as computed by the Entropic Signal Processing System formant tracking algorithm discussed in section 1.2.2.

The order of LPC analysis is based on speaker gender. It is generally agreed that LPC analysis is most useful if one complex pole pair is allocated to each expected formant, plus two complex pole pairs to represent variations in the spectral tilt. In the Entropic formant tracker, speech is downsampled to a 10000 Hertz sampling rate before LPC analysis. Males and females are expected to have five and four formants, respectively, in the first 5000 Hertz, so male voices were analyzed using fourteenth order LPC analysis (7 complex pole pairs), and females were analyzed using twelfth order analysis.

3.1.2 Time-averaged power spectrum of the burst

During transient and frication excitation, the transfer function is not well modeled by an all-pole model. The speech production theory sketched in chapter 2 suggests that measurements of the frequency and amplitude of the front cavity resonance, and of the number of measurable back cavity resonances in the spectrum, may be useful for classification.

In this chapter, burst spectral information is measured from a time-averaged power spectrum $P_x(f)$, computed using the algorithm discussed in section 2.1.4. DFT spectra are first computed using 6ms Hamming windows, and with zero preemphasis. DFT amplitudes are squared to provide an estimate of the power spectrum, and 7-10 consecutive overlapping power spectra are averaged to form a smoothed spectral estimate (step 1ms; number of spectra to average depends on the experiment). The Hamming windows are located as late as possible in the signal, provided that the first window is centered on or before the transcribed release, and the last window is centered no less than 2ms prior to the transcribed voice onset.

3.2 Knowledge Representation and Algorithm Design

This section describes the knowledge-based development of algorithms for accurate measurement of formants during aspiration and voicing, and of front cavity resonance information during frication. In this section, the prior knowledge of a human transcriber is formalized using Bayesian *a priori* distributions. These *a priori* distributions, combined with some knowledge of the randomness in the signal representation, are used to derive measurement algorithms which maximize an *a posteriori* probability of correctness. It should be noted that human judges are more likely to think in terms of measurement algorithms and rules than they are to think in terms of Bayesian priors; the Bayesian analysis pursued here is merely an attempt to formalize the knowledge representation.

3.2.1 Burst front cavity resonance

Both speech production theory and articulatory measurements suggest that the place of articulation of a stop is almost uniquely specified by the length of the front cavity. In almost all cases, the first front cavity resonance frequency F_{f1} is related to the front cavity length l_f by the simple formula $F_{f1} = c/4l_f$, so that a correct measurement of the first front cavity resonance frequency almost always specifies place of articulation. This section presents

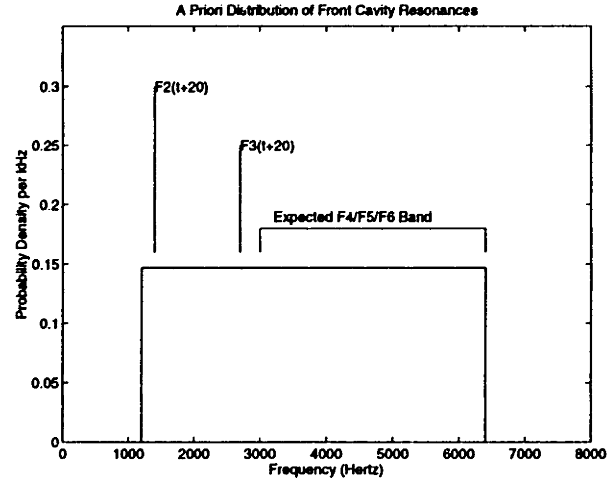


Figure 3-1: Schematized *a priori* distribution of possible front cavity resonance frequencies. F2(t+20) and F3(t+20) are the F2 and F3 measurements 20 milliseconds after release, which are assumed to be the first reliable formant measurements. With no knowledge of the stop or vowel identity, the most we can say is that the front cavity resonance must be between the vicinity of F2(t+20) and some fixed upper bound.

Bayesian models of theoretical and empirical knowledge about front cavity resonance peaks, and shows how the models can be used to specify a measurement algorithm.

A priori distribution

Prior studies indicate that the front cavity resonance of a velar stop usually equals the onset frequency of F2 or F3 of the following vowel, while the front cavity resonance of an alveolar stop usually equals the onset frequency of F4, F5, or perhaps F6 (e.g. Stevens, 1996). With no other *a priori* information about the placement of the front cavity resonance, it seems reasonable to assume that the front cavity resonance is equally likely to be at any of these formant frequencies.

Unfortunately, formant measurements at release are unreliable, and an estimate of the front cavity resonance range based on erroneous formant measurements might, in the worst case, not include the real resonance peak. There are two solutions: we can use measurements from aspiration or voicing (at least 20 ms after release), or we can use a constant frequency threshold representing the most extreme expected formant.

Measurements of F4, F5, and F6 are likely to be erroneous, even if measured in clear voicing, so it is probably most useful to assume that the front cavity resonance of an alveolar is equally likely to take any frequency below some reasonable constant upper bound. The value of this upper bound will be determined empirically in section 3.3 to be roughly 6300 Hertz.

F2 and F3 can usually be measured with some accuracy within 20-30 milliseconds after release, but the formant frequencies 20-30 milliseconds after release may be different from the onset frequencies at the moment of release. In order to make use of a delayed measurement of F2, we will have to make allowance for a reasonable amount of formant motion at onset.

This chapter assumes the simplified *a priori* distribution shown in figure 3-1. In this distribution, the front cavity resonance is assumed to be equally likely to take any frequency between the F2 onset region and about 6300 Hertz. The F2 onset region is defined as the set of frequencies within some maximum distance of a post-frication F2 measurement. This maximum distance will be determined empirically in section 3.3 to be roughly 200 Hertz.

Variance in the signal representation

In looking for front cavity resonances, we begin with the assumption that a front cavity resonance always appears as a convex peak in the DFT spectrum. This assumption greatly simplifies analysis, and is almost always true. According to human judges, the front cavity resonance was marked by a peak in all of the 84 alveolar and velar training tokens, and 143 of the 144 alveolar and velar test tokens.

According to the model, then, if there is a front cavity resonance, there are always one or more convex peaks in the burst spectrum between the F2 onset frequency and 6000 Hertz, one of which must be associated with the front cavity resonance. Of the other peaks, some may be caused by random fluctuations in the source spectrum, some may be caused by back cavity or constriction resonances, and, if the first front cavity resonance is below 2000 Hertz, one may be the second front cavity resonance. If there is no front cavity resonance (that is, if the stop is labial), there may be peaks caused by back cavity resonances and random fluctuations, or there may be no clearly defined spectral peaks.

The amplitude distributions of front cavity resonances, back cavity resonances, and random spectral fluctuations were explored in section 2.2.6. In that section, we concluded that the first front cavity resonance is almost always the largest peak in a burst spectrum, with the possible exception of some high-frequency alveolar resonances. In a Bayesian analysis, the dominance of the front cavity resonance can be represented by a model in which the probability distribution of front cavity resonance amplitudes is a monotonically increasing function of amplitude. The simplest such model, assuming we begin with a log-magnitude spectrum, is one in which the distribution of front cavity resonance amplitudes, $P_x(\widehat{F_{f1}})$, is logarithmic between an arbitrary minimum amplitude and an arbitrary maximum amplitude, independent of the amplitudes of any other peak in the spectrum:

$$\Pr(P_x(\widehat{F_{f1}})) \propto \log\left(\frac{P_x(F_{f1})}{X_{MIN}}\right), \quad X_{MIN} < P_x(F_{f1}) < X_{MAX} \qquad (3.1)$$

The theory in section 2.2.6 suggests that equation 3.1 should be modified slightly, to represent the negative correlation between amplitude $P_x(\widehat{F_{f1}})$ and frequency F_{f1}. Section 3.3 will describe a simple empirical rule, in which the amplitude of a high-frequency front cavity resonance is occasionally as much as 1dB lower than the amplitudes of lower-frequency back-cavity resonance peaks in the same spectrum.

Algorithm design

Given the amplitudes and frequencies of convex peaks in a burst spectrum, figure 3-1 and equation 3.1 can be combined to determine the *a posteriori* probability distribution of the front cavity resonance. The resulting distribution is shown in figure 3-2.

According to the model, the probability of the front cavity resonance equaling any frequency which is not a convex peak, or which is outside of the range shown in figure 3-1, is zero. What remains is a discrete set of frequencies, corresponding to the convex spectral peaks F_{cn} in the range of interest. The *a posteriori* probability of the front cavity resonance

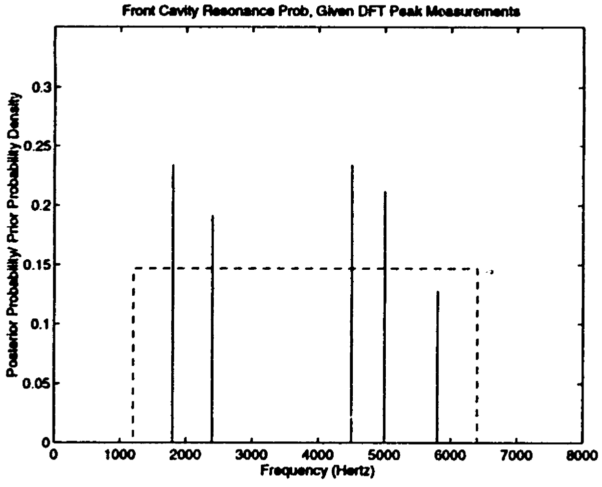


Figure 3-2: Schematized *a posteriori* probability distribution of the front cavity resonance frequency F_{f1}, given a particular burst spectrum. In the example shown, five convex spectral peaks have been identified in the frequency range of interest, with log amplitudes of 74, 60, 74, 67, and 40dB, respectively (measured relative to an arbitrary minimum amplitude). According to formulas specified in the text, the *a posteriori* probability of the front cavity resonance being located at any of these five peak frequencies is proportional to the log amplitude of the corresponding spectral peak.

equaling any peak in this set is proportional to the log amplitude of the peak:

$$\Pr(F_{f1} = F_{cn} | P_x(\widehat{F_{c1}}), ..., P_x(\widehat{F_{cN}})) = \frac{\log(P_x(\widehat{F_{cn}}))}{\sum_{i=1}^{N} \log(P_x(\widehat{F_{ci}}))} \tag{3.2}$$

A maximum likelihood estimation algorithm based on equation 3.2 simply picks the largest spectral peak in the range of interest. If two peaks have nearly the same amplitude, further information is needed to differentiate the two. Empirical investigations in section 3.3 will suggest that, for the particular microphone and analysis conditions used here, if two peak amplitudes differ by less than 1dB, the peak which is higher in frequency should be preferred.

3.2.2 Other burst spectral measurements

Other burst spectral measurements may also be useful in distinguishing between the three stop places of articulation. This section briefly discusses measurements of burst spectral amplitude, and of the number of back cavity resonances.

In addition to the front cavity resonance measurement described above, there should also be some way of identifying bursts which have no front cavity resonance, since in these cases, the peak-finding algorithm described above produces a meaningless answer. According to the production theory in section 2.3, the front cavity resonance of an alveolar or velar stop gives the burst spectrum a 10-20dB boost at the resonant frequency. This implies that velar stops are 10-20dB more intense than labial stops in the F2-F3 region of the spectrum, and alveolar stops are more intense than labials in the F4-F5-F6 region of the spectrum (provided that the recording levels are similar). Amplitude measurements in these two frequency ranges are therefore used to discriminate labial stops from each of the other two places of articulation, with the exact frequency ranges determined empirically in section 3.3.

According to section 2.2.2, stops which are released quickly have a lower constriction impedance, and therefore more coupling between the front and back cavities, than stops which are released slowly. Section 2.3 suggests that labial and alveolar stops are released more quickly than velars. To the extent that this is true, there should be more back cavity resonances visible as pole-zero pairs in the burst spectra of labials and alveolars than in the burst spectrum of a velar. If this distinction exists, it should be possible to take advantage of it by simply counting the number of significant peaks in the burst spectrum, where the definition of "significant" will be determined empirically in section 3.3.

3.2.3 Formant frequencies above F1

The roots of a carefully measured LPC polynomial provide a good estimate of formant frequencies during voicing, and an LPC root-finding algorithm, such as that used by the Entropic formant tracking algorithm, is a good place to start when looking for formants. During aspiration, however, LPC may occasionally mistakenly find a tracheal resonance instead of a formant, while during frication, production theory predicts that only front cavity resonances show up as roots of the LPC polynomial. Human transcribers are very good at making use of their knowledge of formant continuity to fill in sections of a formant track which are missing from the signal. This section presents a Bayesian model of human knowledge about formant tracking, and then apply the model to derive an algorithm for smoothing the poles of an LPC formant tracker.

A priori distribution

Formants are the natural frequencies of a given physical system (the vocal tract, between the glottis and the lips), and as such, they are constrained to change continuously as a function of time, except occasionally when a formant crosses the frequency of a subglottal resonance (see section 2.2.5). We can model this constraint by modeling the *a priori* likelihood of a formant's location at each time as a bell-shaped curve, centered on the formant value which would be predicted using continuity constraints.

Figure 3-3 shows a simple *a priori* distribution, in which the formant prior at time t is a function only of the measured formant at time $t + \Delta t$, for some suitable step size Δt. The width of the bell-shaped curve should be determined empirically, but it should also be made large enough to allow for occasional discontinuities of 200-300 Hertz as a formant crosses the frequency of a subglottal resonance. Section 3.3 will determine a general width parameter which is a function of the width of this curve.

Signal variance

Careful LPC analysis can generally be assumed to produce accurate formant measurements during modal voicing, that is, beginning one or two pitch periods after the transcribed voice onset. LPC analysis of an aspiration spectrum produces occasional errors, as discussed in section 2.2.5. LPC analysis of a frication spectrum, during the first 20 milliseconds or so following a stop release, is generally expected to measure only formants which happen to be associated with the front cavity, and to largely ignore back cavity resonances.

Formant tracking errors can be small errors, in which the frequency of a peak is shifted slightly because of a random spectral null, or large errors, in which the LPC polynomial models entirely the wrong peak. Formally, we can assume that measurement errors ϵ are

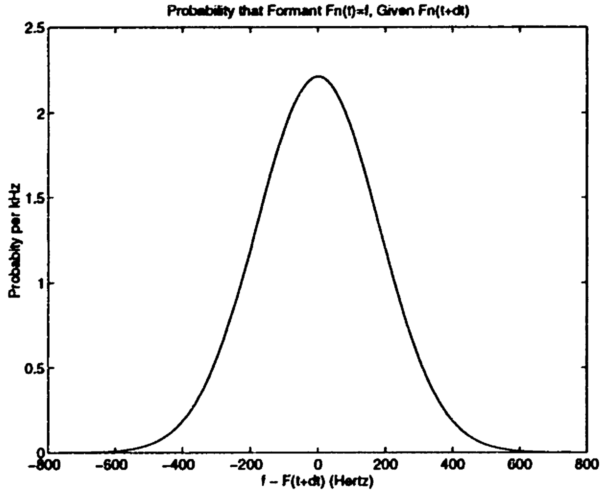


Figure 3-3: Schematized *a priori* distribution of a given formant frequency, as a function of a later measurement of the same formant. With no information about the consonant or the vowel, we can only predict that the formant is continuous, hence a bell-shaped distribution.

normally distributed over a narrow region with probability $1-P$, and normally distributed over a much broader region with probability P:

$$P(\epsilon) = (1-P)N(\frac{\epsilon}{\sigma_1}) + PN(\frac{\epsilon}{\sigma_2}), \quad \sigma_2 \gg \sigma_1 \tag{3.3}$$

where $N(x)$ is the unit normal distribution.

Our knowledge that formant errors are more likely during frication than during voicing is easily incorporated into equation 3.3 by making P a function of time. The likelihood of formant errors during frication can be represented by a large P during the first 20ms after release. Formant errors during aspiration are less likely, so P can be slightly lower during aspiration. Finally, if formant errors during modal voicing are judged to be impossible, P can be set to zero beginning 10-20ms after the onset of voicing.

Measurement design

Equation 3.3 and figure 3-3 can be combined to specify an algorithm which smooths each formant track backward from the vowel toward consonant release. The empirical study discussed in section 3.3 will determine that LPC formant measurements after about 20ms of voicing can be considered to be reliable. Working backward from the vowel, the *a posteriori* probability distribution of a formant at time t, given the formant measurement at time $t+\Delta t$, can be calculated by multiplying the error model of equation 3.3 by the continuity model shown in figure 3-3.

Figure 3-4 shows two examples of *a posteriori* distributions which might be used in the extension of a formant $Fn(t+\Delta t)$ backward in time to $Fn(t)$. In the top part of the figure, the LPC polynomial at time t has a root which is close to $Fn(t+\Delta t)$, and this root is therefore judged to be the most likely extension of the formant. In the bottom part of the figure, the closest LPC root at time t is far from $Fn(t+\Delta t)$, and is therefore judged to be an unlikely extension of the formant. Since the LPC polynomial provides no other formant candidates, the maximum likelihood estimate of the formant is just $Fn(t) = Fn(t+\Delta t)$.

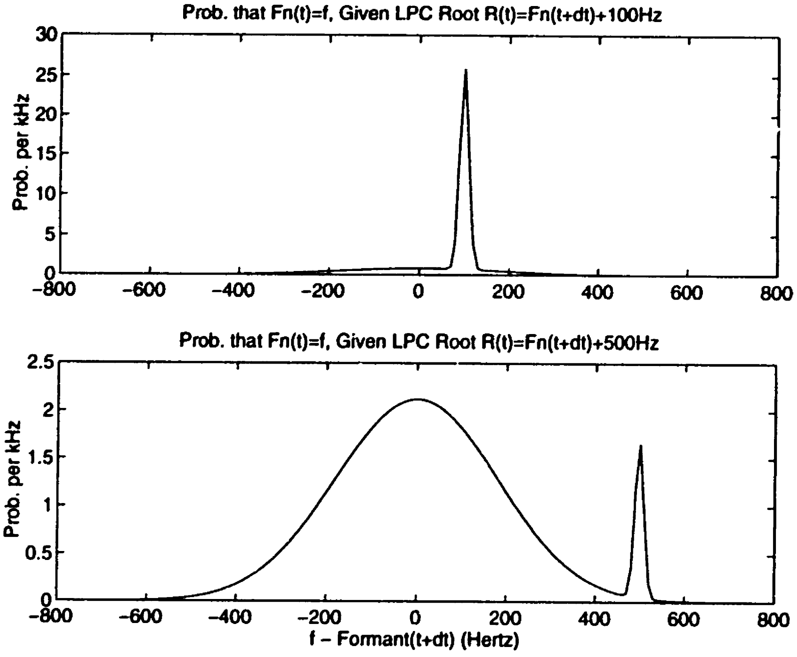


Figure 3-4: Schematized *a posteriori* probability distributions for two formant frequency tracks. In the upper plot, the LPC root at time t is close to the formant at $t + \Delta t$, and is therefore judged to be a continuation of the formant. In the lower plot, the closest LPC root at time t is far from the formant at time $t + \Delta t$, and is therefore judged to be a measurement error.

3.2.4 First formant measure

The first formant is subject to the same continuity constraints as all of the higher formants, but the first formant must also satisfy one additional constraint: at the release of a consonant, the frequency of the first formant should always rise. Transient ringing and subglottal resonances at release of a stop, and nasal resonances at release of a nasal, often introduce strong peaks between roughly 600 and 1000 Hertz which hide the rise of the first vocal tract formant.

Figure 3-5 shows an *a priori* distribution of $F1(t)$, as a function of $F1(t + \Delta t)$, which represents both the continuity of F1 and the fact that F1 is expected to rise as a function of t. The distribution is bell-shaped, but asymmetric. F1 is expected to rise: the expected value of $F1(t)$ is lower than $F1(t+\Delta t)$. Given no other information, however, the maximum likelihood estimate of F1 at time t is still $F1(t) = F1(t + \Delta t)$.

The model of measurement uncertainty for F1 should be the same as it is for other formants, although the probability of an erroneous measurement P may be higher for F1 during aspiration than it is for other formants.

If $P = 0$ during modal voicing (i.e. beginning 10-20ms after voice onset), F1 can be smoothed backward in time from the vowel, just like the other formants. Tracing backward into the release, F1 at time $t + \Delta t$ is connected backward to the lowest root $F_{r1}(t)$ of the LPC polynomial at time t. If $F_{r1}(t)$ is lower than $F1(t + \Delta t)$, as shown in figure 3-6a, then F_{r1} is usually the most probable formant at time t. If $F_{r1}(t)$ is higher than $F1(t + \Delta t)$ by more than some threshold value, $F_{r1}(t)$ is judged to be a subglottal or nasal resonance, and the maximum likelihood estimate of the formant is $F1(t) = F1(t + \Delta t)$.

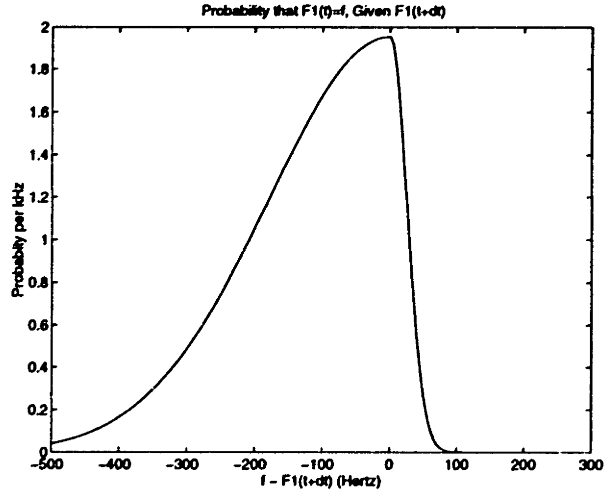


Figure 3-5: Schematized *a priori* distribution of $F1$ at time t, given a measurement of $F1(t+\Delta t)$. Distribution is skewed, representing our expectation that F1 rises as a function of t.

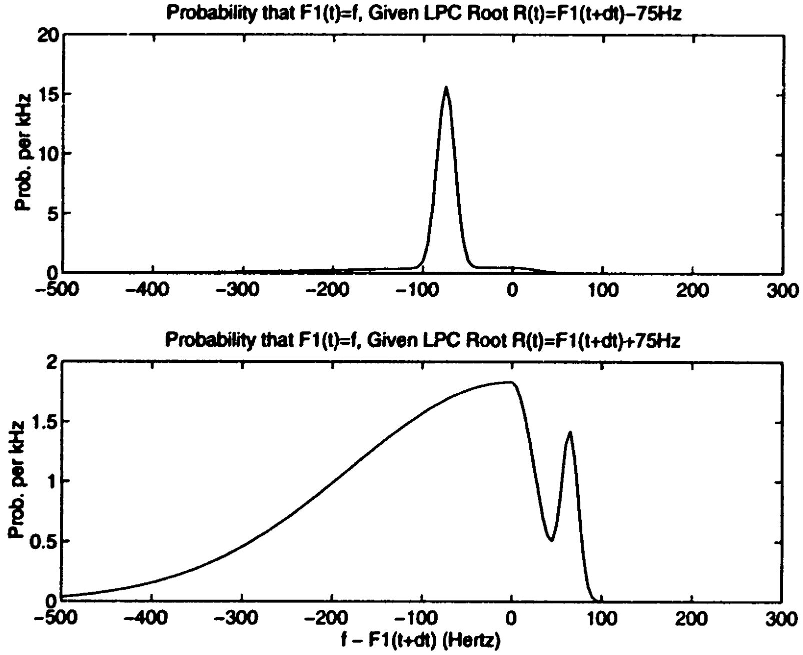


Figure 3-6: Schematized *a posteriori* distributions of F1 at time t, given an LPC root $F_{r1}(t)$ which is (a) 75 Hertz below $F1(t+\Delta t)$, or (b) 75 Hertz above $F1(t+\Delta t)$.

3.3 Imitating Human Performance on a Training Set

The presentation in section 3.2 describes the algorithm design process as an exercise in the application of speech production theory. In fact, speech production theory allows many variations on each of the described algorithms. This section describes experiments in which the precise forms of the algorithms described above, and the values of their temporal and frequency thresholds, are adjusted by a human judge to meet an empirical performance criterion.

3.3.1 Training data

The algorithms outlined in section 3.2 were adjusted by a human judge (the author of this thesis) so that the algorithms would imitate his measurements on a training set of consonant releases, with as few large measurement errors as possible. The training set consisted of 20 tokens of each consonant, split evenly by speaker gender, with right contexts drawn at random from the vowels, glides, and liquids in TIMIT (including 11 alveolar tokens in retroflex context, but none in lateral context). This database is referred to in this thesis as the KB Train (knowledge-based training) database; a list of tokens is provided in appendix A, section A.1.

The judge attempted to produce measurements of the true vocal tract resonances 20ms and 50ms after consonant release, and of the true front cavity resonance at the instant of release. In addition to measurements of the vocal tract resonances, two burst spectral amplitudes and a convex peak count were measured. Measurements of low-frequency and high-frequency spectral amplitude are designed to represent, as much as possible, the amplitudes of front cavity resonances in velar and alveolar stop spectra, respectively. The convex peak count is defined as the number of convex peaks in the spectrum within some threshold distance of the spectral maximum; the amplitude threshold was adjusted to represent, as accurately as possible, the distinction between "compact" velar spectra and "diffuse" alveolar spectra.

All measurements were performed non-interactively, using a spectral representation similar to that available to the automatic measurement algorithms. Burst spectral information was measured from a single time-averaged power spectrum, consisting of the average of seven consecutive squared DFT spectra (step 1ms). Formant frequency measurements were based on a list of candidate formants generated by the ESPS formant tracker, combined with time-averaged power spectra 20ms and 50ms after consonant release. The judge was given the identity of each consonant, and its phonetic context.

3.3.2 Formant measurement algorithms

An algorithm for automatic measurement of onset and vowel target formants is defined by the application of simple smoothing rules. All automatic formant measurements are based on roots of the LPC polynomial, as calculated by the formant tracker packaged with the Entropic Signal Processing System. No gross formant measurement errors were found in the training data more than 10ms after the transcribed voice onset, so formant tracks proposed by the ESPS tracker are assumed to be 100% correct beginning 20ms after the transcribed voice onset. Earlier formants are smoothed backward from the vowel toward the release, with a 5ms step between analysis frames.

First formant measures

At each time t, the first formant $F1(t)$ is estimated on the basis of the first root of the LPC polynomial, $F_{r1}(t)$, and of the known first formant at time $t + 5$ms, $F1(t+5)$. If $F_{r1}(t)$ is too high, it is judged to be a subglottal or nasal resonance, and the formant estimate is $F1(t) = F1(t+5)$. If $F_{r1}(t)$ is sufficiently low, it is judged to be a continuation of the formant track: $F1(t) = F_{r1}(t)$. A threshold specifying that $F_{r1}(t) \leq F_{r1}(t+5) + 50$ Hertz was determined to divide these two cases for almost all training tokens.

Second and third formant measures

At each time t, the nth formant $Fn(t)$ is estimated on the basis of its known value at time $t + 5$ms, $Fn(t+5)$, and of the root of the LPC polynomial $F_{rj}(t)$ closest in frequency to $Fn(t+5)$. The formant track $Fn(t)$ must be allowed to jump discontinuously across the frequency of a subglottal pole, but should otherwise be constrained to be as continuous as possible. In the KB Train database, a small number of measured formant tracks were discontinuous by more than 300Hz in 5ms, but none were discontinuous by as much as 400Hz, so a rule was implemented requiring formant frequency discontinuities to be less than 400Hz. If the absolute difference between $Fn(t+5)$ and $F_{rj}(t)$ is less than 400 Hertz, then $Fn(t) = F_{rj}(t)$, else $Fn(t) = Fn(t+5)$.

3.3.3 Burst spectral measurement algorithms

All burst spectral measurements depend on the definition of a frequency band of interest, corresponding to the frequency band in figure 3-1 in which the front cavity resonance has a non-zero *a priori* probability of occurrence. This band-limited spectrum will be referred to as the front cavity resonance spectrum.

The front cavity resonance spectrum, and the four burst spectral measurements derived from it, are defined by a small number of rules. Application of these rules to the convex peaks of the burst spectrum produced measurements similar to those transcribed by the human judge.

Front cavity resonance spectrum

The front cavity resonance spectrum is defined to be a band-limited portion of the time-averaged power spectrum of the burst, where the power spectrum was computed, in this experiment, as the average of 7 consecutive squared DFT spectra (step 1ms), with zero preemphasis.

A high-frequency boundary at 6300 Hertz was found to be high enough to include all alveolar front cavity resonances in the training data. The low-frequency boundary was defined as being 200 Hertz below a measurement of F2. The measurement of F2 20ms after release was determined to be early enough to reliably indicate the frequency of a velar front cavity resonance, but late enough to be rarely influenced by frication noise.

Front cavity resonance frequency

The highest-amplitude peak in the front cavity resonance spectrum was found to correspond to the front cavity resonance in almost all cases.

In three cases, the front cavity resonance spectrum of a non-retroflex alveolar stop contained a back cavity resonance, at the frequency of F2, with an amplitude equal to or

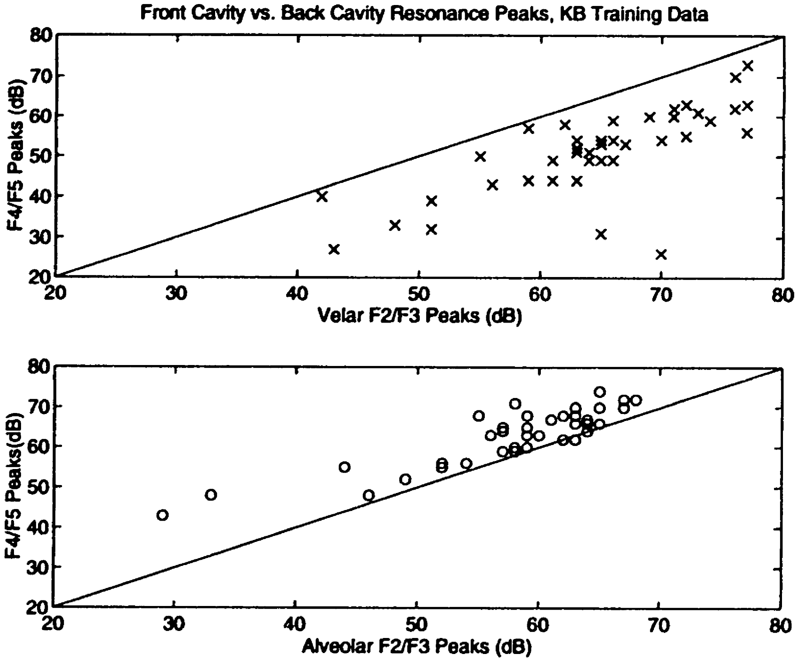


Figure 3-7: Amplitudes of the highest peak in the F2/F3 and F4/F5/F6 range, velar (top) and non-retroflex alveolar (bottom) tokens only.

1dB higher than that of the front cavity resonance (figure 3-7). In order to correctly identify the front cavity resonances of these three training tokens, a rule was implemented favoring the highest in frequency out of any set of peaks with amplitudes within 1dB of each other.

In the burst spectra of the 11 retroflex alveolar stops in the KB Train database (not shown in the figure), the front cavity resonance frequency was always the largest peak in the spectrum, with no peaks of similar amplitude at higher frequency. All but one of the 11 retroflex stops had front cavity resonance frequencies between 2000 and 3500 Hertz.

Number of peaks

As a measure of compactness, the number of "significant" peaks in the front cavity resonance spectrum is counted. A definition of "significance" in terms of relative amplitude was found to provide a reasonable measure of compactness: a "significant" peak is defined to be a peak whose amplitude is no more than 10dB below the amplitude of the largest peak in the front cavity resonance spectrum.

Amplitude measures

Burst amplitude measurements were designed to discriminate between labial stops, which do not have a front cavity resonance peak, and alveolar and velar stops, which do. It was discovered that, in the KB Train database, the best discrimination between these two classes is provided by a pair of frequency bands which, combined, cover a range slightly larger than the front cavity resonance spectrum. Low-frequency amplitude is measured to be the amplitude of the highest peak in the band between 1000 and 3400 Hertz, and therefore often includes strong back cavity resonances at the release of an alveolar stop. The high-frequency amplitude is measured to be the peak amplitude in the band between

2700 Hertz and 7400 Hertz.

If there is no peak in a band, it is not clear how the "peak amplitude" should be defined. In classification experiments, it is useful to assign bands without peaks a "peak amplitude" measure which is lower than any peak amplitude observed in the same band in tokens which do have peaks, but not too much lower. In experiments in this thesis, two alternate and equally arbitrary conventions were adopted: the "peak amplitude" measure of a band with no peaks was sometimes set to 15dB below the overall DFT peak amplitude, and was sometimes set to the amplitude of the lowest in-band spectral valley. Neither of these performed better than the other in classification experiments.

3.4 Statistical Models of Measurement Error

An automatic measurement algorithm is not useful for phonetic studies without reliable estimates of the aggregate measurement error. This section describes experiments in which the performance of the algorithms described above was compared to the transcriptions of human judges on a test set. The distribution of measurement errors was then modeled using several statistical models, and confidence limits on the error were computed.

3.4.1 Reference measurements

The algorithms described in section 3.3 were tested on a database consisting of 324 consonant releases: nine consonants (six stops, three nasals) × two genders × eighteen right contexts. This database is referred to in this thesis as the Error Modeling database; a list of tokens is given in appendix A, section A.2.

Two human judges, each with at least five years of phonetic training, attempted to measure vocal tract resonances at the given stop releases (one of the judges was the author of this thesis). Judges were given full information about each sentence, including the transcription, and were allowed to use any spectral representations which they found useful in estimating the requested measurements.

The judges were instructed to measure vocal tract resonance frequencies as accurately as possible at six specified times in each waveform (at ten millisecond intervals, from 5 ms to 55 ms after the transcribed release). One of the judges relied primarily on DFT spectra computed with a 14 ms Hamming window, supplemented by LPC spectra computed with a 20 ms window; the other judge chose to use time-averaged DFT spectra with a 10 millisecond averaging window.

In addition to the vocal tract resonance measurements, judges were asked to make several measurements on the burst spectra of the stop releases. First, for velar and alveolar stops, judges were asked to measure the first front cavity resonance frequency. Second, judges were asked to measure the amplitudes of the highest peaks in two bands, which were defined in terms of formants: the "low-frequency" band was defined as the band containing the speaker's typical F2 and F3, while the "high-frequency" band contained the speaker's typical F4, F5, and F6. Finally, the judges were asked to estimate the "diffuseness" of the spectrum, on a scale from 1 (most compact) to 5 (most diffuse).

Roughly one-third of the consonants were transcribed by both judges. After transcribing these syllables individually, the judges compared their measurements to make sure that they were choosing the same peak for each frequency measurement. Remaining differences between the judges should therefore be entirely caused by differences in the signal representations used.

Measure	N	r_q	Mean	Std Dev	Min	Max
F1 (Hertz)	561	0.99	-13	36	-158	100
F2 (Hertz)	700	0.98	1	47	-257	237
F3 (Hertz)	662	0.99	6	55	-220	225
Resonance (Hz)	63	0.97	10	50	-170	131
Low Freq Amp (dB)	86	0.97	16.0	3.2	3.7	27.0
High Freq Amp (dB)	87	0.96	15.5	3.1	6.2	21.5
Diffuseness	88	0.93	0.3	1.0	-2	3

Table 3.1: Signed differences between the two human judges on tokens which were transcribed by both. F1, F2, and F3 include six measurements per token, for more than 100 tokens. The burst spectral measures are the front cavity resonance frequency, low frequency and high frequency amplitudes, and relative diffuseness.

The difference between the judges' scores on each measurement are presented in table 3.1. In this table, r_q is a test of the goodness of fit of a Gaussian model, which will be described in the next subsection. The high values of r_q in table 3.1 indicate that all differences are relatively well modeled by a Gaussian distribution, except that differences in the diffuseness measure are always small integer values (usually 0 or 1), and are therefore not well modeled by a Gaussian. Since each judge was allowed to choose his own spectral representation, there is a large mean difference in the amplitude measurements.

3.4.2 Confidence limits for a Gaussian error model

Measurements of the two judges were combined to form a reference set, against which the algorithm was tested. For all tokens transcribed by both judges, the two sets of measurements were averaged. Amplitude measurements of one judge were linearly shifted by the amount of the average difference before averaging; frequency and diffuseness measurements were averaged with no prior adjustment.

The algorithms described previously were used to automatically measure formant frequencies, burst front cavity resonance and amplitudes, and diffuseness (measured as the number of large peaks in the burst spectrum). These measurements were compared to the reference measurements produced by human judges, and the difference will be referred to as the "measurement error" of the algorithm.

Table 3.2 gives estimates, and 99% confidence limits, for the error mean and error standard deviation of each measurement. In this table, the large average difference between automatic and manual measurements of spectral amplitude is caused by the different spectral representations used by humans and machine. There are also significant differences between the automatic and manual measurements of F1, F3, and the front cavity resonance, which can not be easily explained in terms of differences in spectral representation.

Differences shown as significant in table 3.2 should be interpreted with caution, because of the low quantile correlation coefficients r_q listed in the table, and the obvious nonlinearity of the Q-Q (quantile-quantile) plots shown in figure 3-8. A Q-Q plot (Johnson and Wichern, 1992) displays the measurement errors, sorted in increasing order, plotted against an equal number of samples drawn uniformly from a Gaussian distribution. If the measurement errors are also drawn from a Gaussian distribution, the Q-Q plot should be a straight line, and the normalized correlation r_q between the ordinate and abscissa should equal 1.0. Instead, all of the r_q coefficients in table 3.2 are significantly less than unity ($p < 0.01$), although errors

Formant Measures

Measure	N	r_q	Mean (99% Limits)	Std Dev (99% Limits)
Onset F1 (Hz)	594	0.91	-43 (-55,-31)	114 (106,123)
Vowel F1 (Hz)	827	0.84	-26 (-34,-18)	86 (81, 92)
Onset F2 (Hz)	871	0.85	-1 (-20, 18)	214 (201,228)
Vowel F2 (Hz)	925	0.73	6 (-10, 22)	188 (177,200)
Onset F3 (Hz)	864	0.91	40 (21, 59)	216 (203,230)
Vowel F3 (Hz)	919	0.78	38 (21, 54)	193 (182,206)

Burst Spectral Measures

Resonance (Hz)	142	0.77	-156 (-304,-8)	674 (584,794)
Low Freq Amp (dB)	214	0.98	-51.5 (-52.8,-50.2)	7.3 (6.5,8.3)
High Freq Amp (dB)	214	0.99	-56.6 (-57.5,-55.6)	5.3 (4.7,6.1)
Diffuseness	214	0.97	-0.2 (-0.5,0.0)	1.5 (1.3,1.7)

Table 3.2: Mean and standard deviation of differences between automatic measurements and measurements transcribed by human judges, with bilateral 99% confidence limits. "Onset F1" includes measurements 0, 10, and 20ms after release, "Vowel F1" includes measurements 30, 40, and 50ms after release.

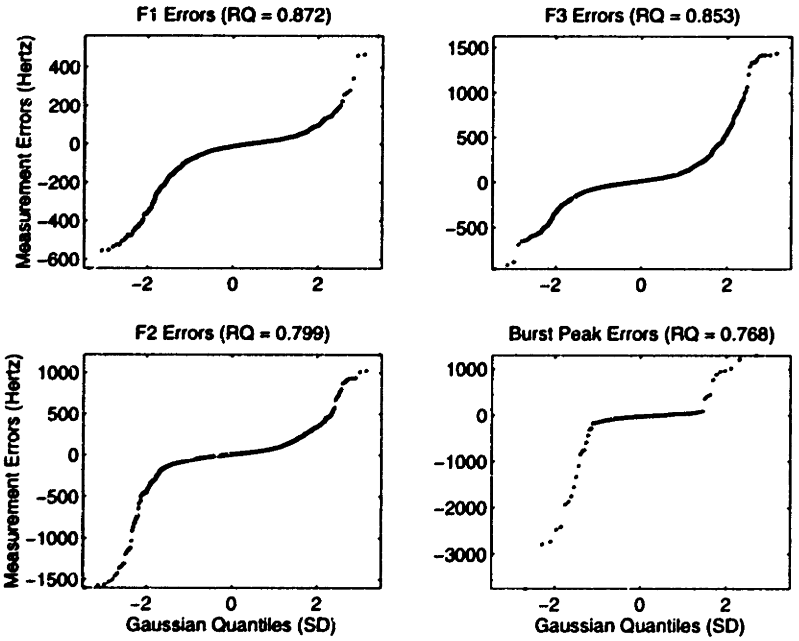


Figure 3-8: Q-Q plots describing the degree to which the measurement error fits a Gaussian model: if the errors fit a Gaussian model, the plot should be a straight line. Low-amplitude errors are approximately Gaussian, but there are many outliers.

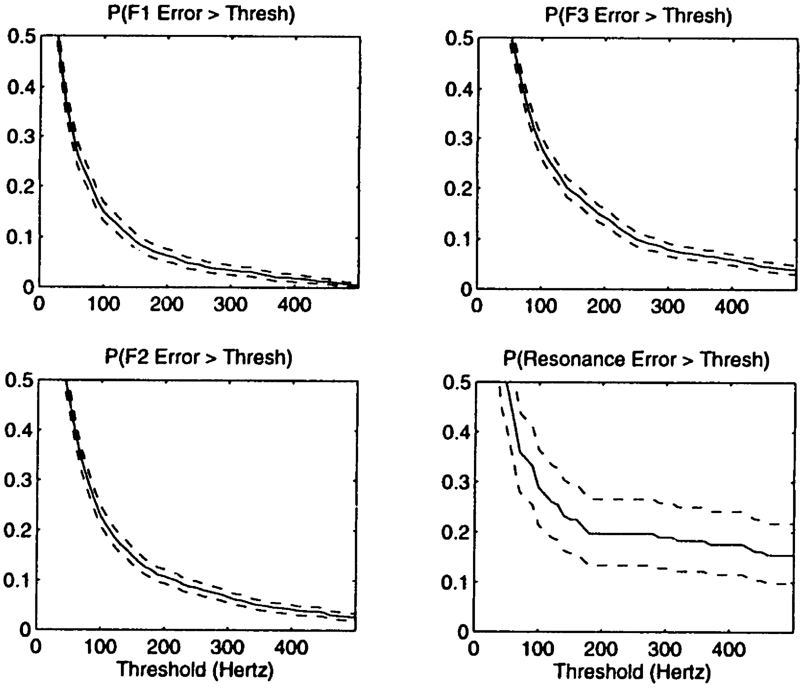


Figure 3-9: The cumulative probability of finding a measurement error greater than any given frequency, with 95% confidence limits. Plots have been truncated at 500 Hertz, in order to make the threshold behavior near 150 Hertz more visible.

in the measurement of spectral amplitudes and number of peaks (not shown) can probably be considered Gaussian for most practical purposes.

The reason that the errors do not fit a Gaussian model is clearly displayed in figure 3-8. In this figure, the smaller errors fit the normal error model quite well, but the larger errors (both positive and negative) are much larger than the errors that would be produced by a Gaussian model.

Measurement errors which are too large to come from a normal distribution are often called outliers, and often result from the amplification of small errors by a nonlinear process. In this case, the process of peak picking is decidedly nonlinear: a slight change in the relative amplitudes of formant and subglottal resonances, for example, can cause LPC to choose the wrong peak. These large outliers are qualitatively different from the low-amplitude errors, which are caused by more normally distributed sources of error, including possibly peak centering strategies, and differences in signal representation.

The next two subsections consider approaches to ιhe modeling of these outliers.

3.4.3 Explicit modeling of outliers

In experimental situations, a useful model of measurement error should include, first, an estimate of the likelihood of the outliers, and second, an estimate of their size. This section considers techniques for explicitly modeling measurement outliers using nonparametric tools and mixture Gaussian distributions.

	Outlier	r_q		
Measure	Threshold	All Errors	Small Errors	Large Errors
F1	121 Hz	0.872	0.991	0.954
F2	156 Hz	0.799	0.997	0.954
F3	169 Hz	0.853	0.997	0.976
Resonance	180 Hz	0.768	0.985	0.979

Table 3.3: Separation of measurement errors into low-amplitude "normal errors" and high-amplitude "outliers." Measurements are F1, F2, and F3 (six measurements per consonant release) and the front cavity resonance (one measurement per stop release).

A non-parametric model

Figure 3-9 shows, for the front cavity resonance and all three formants, the probability of encountering an error larger than the given threshold. 95% confidence limits for each measurement are shown using dotted lines. The confidence limits were calculated to three decimal places by iteratively testing the parameters of a binomial distribution. Notice that, for each measurement, the probability of an error drops off steeply until a sort of threshold is reached, and drops off more slowly thereafter. For practical purposes, errors above this threshold can be considered outliers, and errors below this threshold can be considered relatively normal.

Table 3.3 lists approximate values of the threshold for each measurement. The three formant thresholds are each calculated as the inverse of the decay constant in an exponential probability model, where the models were estimated by fitting straight lines to the logarithms of the curves shown in figure 3-9. The exponential probability model does a poor job of modeling details of the distributions, but the estimates of the thresholds for formant outliers were considered reasonable, and are therefore used in table 3.3. The front cavity resonance threshold, on the other hand, was read directly from figure 3-9: the threshold, 180 Hertz, is the start of the long flat section in figure 3-9, and seems to represent a natural boundary in that distribution.

As a measure of the normality of each subset of the errors, table 3.3 lists Gaussian quantile correlations r_q separately for both low-amplitude and high-amplitude errors. Our previous characterization of the low-amplitude errors as "normal errors" seems to be justified by the r_q statistics, which rise when high-amplitude errors are removed. Surprisingly, the degree of normality also seems to rise when the *low* amplitude errors are removed. This seems to imply that the "outliers" may also be normally distributed, but with a much larger variance than that of the low-amplitude errors.

A concentric mixture Gaussian model

The idea that low-amplitude and high-amplitude errors are independently Gaussian, with similar means but different variances, can be modeled using mixture Gaussian models. We can define a "concentric mixture Gaussian model" to be a model composed of two Gaussians with the same mean:

$$p(\epsilon) = (1 - P) \times N\left(\frac{\epsilon - \mu}{\sigma_1}\right) + P \times N\left(\frac{\epsilon - \mu}{\sigma_2}\right) \qquad (3.4)$$

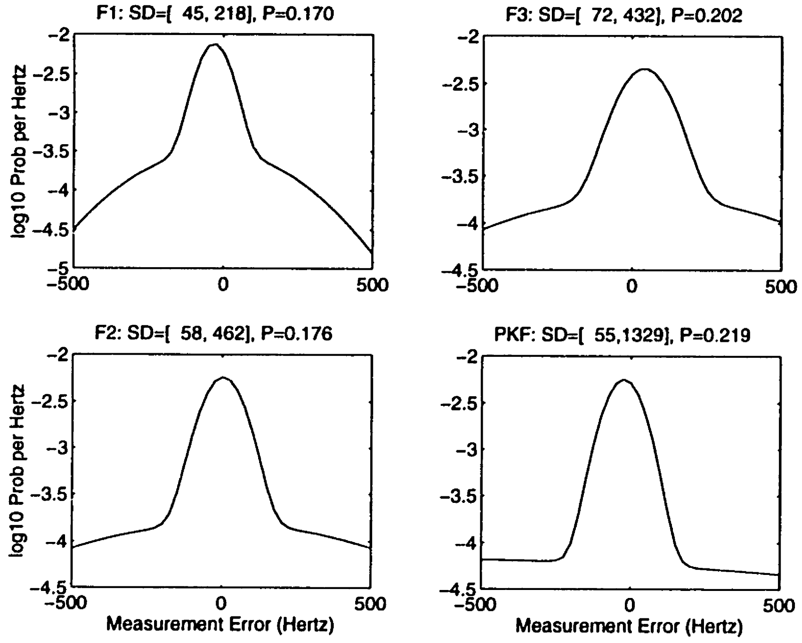


Figure 3-10: Mixture Gaussian models of the formant and front cavity resonance measurement errors (shown: log probability density). Formant mixtures were constrained to have the same mean; the front cavity resonance model was not so constrained.

where $N(x)$ is the unit normal distribution, ϵ is the error, and μ, σ_1, σ_2, and P are trainable parameters.

Concentric mixture Gaussian models trained for the three formant measurement errors are shown in figure 3-10. The front cavity resonance is very poorly modeled by a concentric mixture Gaussian, so the means of the two Gaussian elements were allowed to diverge, creating the slightly skewed distribution shown in the figure.

The parameter P, for each mixture Gaussian model, is slightly higher than the corresponding probability of an outlier shown in figure 3-9. This is a natural consequence of the form of the mixture Gaussian model, which explains almost all outliers using the wide Gaussian element, but adds both Gaussian elements together to explain tokens toward the center of the distribution.

3.4.4 The effect of context on outlier probability

The test set contains nine consonants (three nasals and six stops), and a wide variety of right contexts, including two glides, three retroflex sounds, two lateral sounds, two types of schwa, and nine unreduced vowels. Speech production theory suggests that some of these syllables have clear, easy to measure formants, and others are subject to more frequent errors.

Contexts likely to cause large measurement errors were identified by visual clustering of the test data, with guidance from speech production theory. The resulting definitions of "high-error" and "low-error" contexts are given in table 3.4.

The best predictor of error in an F1 measurement is found to be identity of the consonant. Nasal and aspirated consonants cause frequent peak-picking errors near the consonant release, while only the aspirated consonants with longer voice onset times (/t/ and /k/) cause a noticeable increase in error rate between 30ms and 50ms after release.

	Time	High-Error	Number of Measurements	
Measurement	after Release	Contexts	High-Error	Low-Error
Vowel F1	30-50ms	/t,k/	130	697
Onset F1	0-20ms	same, plus /m,n,ng,p/	379	215
Vowel F2	30-50ms	/w,r,l,y/	209	716
Onset F2	0-20ms	same, plus /g,p,t,k/	503	368
Vowel F3	30-50ms	/w,r,er,axr,l,el,y/	357	562
Onset F3	0-20ms	same, plus /g,p,t,k/	571	293
Peak Frequency	0	Velar + /y/ Alveolar + /w,aa,ae,r,er/	27	115

Table 3.4: Definitions of seven frequency measurements, and of contexts in which the measurements are more likely to suffer peak-picking errors. High-error contexts were chosen on the basis of speech production theory, and of observation of the measured errors.

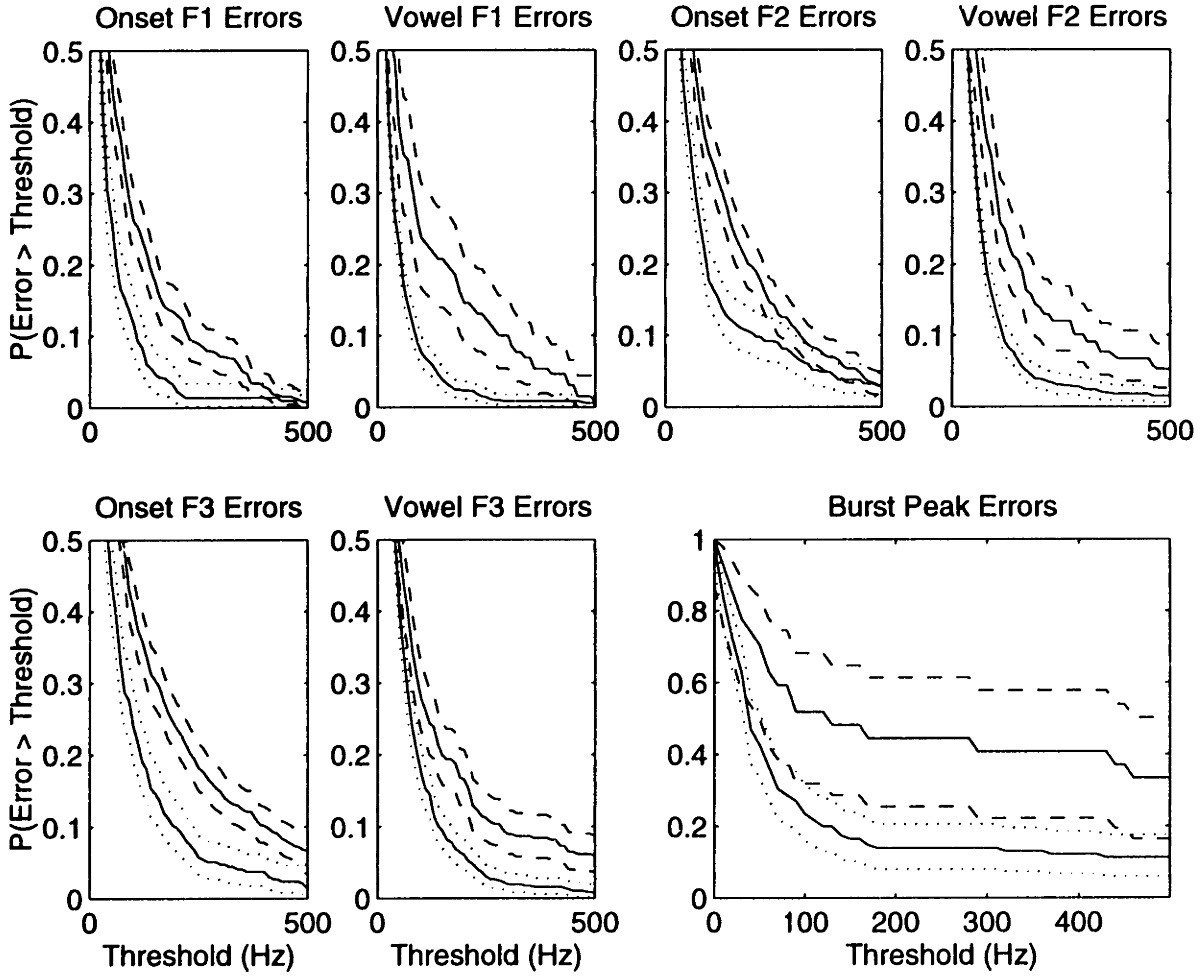


Figure 3-11: Cumulative histograms, showing the probability of finding an error larger than the abscissa coordinate. The solid lines are the measured probabilities for low-error and high-error contexts; the dotted and dashed lines show 95% confidence limits for the low-error and high-error contexts, respectively.

Figure 3-11 shows the probability of finding outliers greater than any given frequency, for both high-error and low-error contexts, with 95% confidence limits. Apparently, most Fl errors larger than about 100 Hertz can be attributed to the effect of the consonant, although a small number of very large errors (400-500 Hertz or more) are also found in low-error contexts.

Errors in F2 and F3 near consonant release are influenced by identity of both the consonant and its right context, but beginning about 30ms after release, it appears that only the right context strongly influences error rate. Both F2 and F3 are frequently mis-identified during glides and liquids, because of frequent mergers between neighboring formants (/y/ and /r/ show frequent F2-F3 mergers, /w/ shows frequent Fl-F2 mergers, and F3 is often weak or ambiguous in /l/). F3 is also often lost during syllabic liquids: during retroflex sounds, F3 often merges with F2, while during lateral sounds, F3 is often extremely weak.

As shown in figure 3-11, the context classes in table 3.4 are relatively effective at predicting the frequency of errors in all F3 measurements, and in F2 measurements at least 30ms after release. The F2 onset measurements in high-error and low-error contexts are not significantly different, indicating that the selected contexts are not a good predictor of F2 onset measurement errors.

The front cavity resonance measurements show a clear division, at about 200 Hertz, between small amplitude "normal errors" and large amplitude outliers. Of the 27 "outliers," 12 occur in the contexts specified in table 3.4. Alveolar bursts in low context (/aa/ and /ae/) often contain strong back cavity resonances which are mistaken for the front cavity resonance. The front cavity resonance of a retroflex stop, and of a velar in /y/ context, often appears as a broad mass of energy composed of several smaller peaks; the measurement algorithm and the human transcribers often choose peak locations which differ by several hundred Hertz. Finally, the burst in the syllable /dw/ is often weak, with no clear front cavity resonance; this was the only context in which voicing of the stop was useful in predicting front cavity resonance measurement errors.

3.4.5 Correlations between measurements and measurement errors

The discussions in sections 3.4.2 to 3.4.4 assume that the error in measuring a formant frequency is uncorrelated with the true underlying value of the formant. This assumption, called the ***homoskedasticity*** assumption, is common in statistical analysis, and it allows us to build powerful error models, as described above. Real-world measurement processes, however, are often ***heteroskedastic***, that is, measurement errors are often correlated with the correct value of the thing being measured (Kennedy, 1992). Section 5.1 will show that, if an error model is to be carried forward into any further analysis of the data, it is important to correctly model the correlation between measurement and error.

If we assume that the measurement and measurement error are approximately Gaussian, it is possible to test the assumption of homoskedasticity using standard linear regression analysis. Let x_0 be a correct formant or burst spectral measurement, x the corresponding automatic measurement, and ϵ the measurement error:

$$x = x_0 + \epsilon \tag{3.5}$$

The joint distribution of normally distributed x_0 and ϵ is completely specified by their means

Normalized correlation between measurement and error

Measure	Correlation (99% Limits)	Measure	Correlation (99% Limits)
Resonance	-0.19 (-0.4,0.0)	F1, 50ms	-0.27 (-0.5,-0.1)
Diffuseness	-0.31 (-0.5,-0.1)	F2, 10ms	-0.23 (-0.5,0.0)
Low-Freq. Amp	-0.55 (-0.7,-0.4)	F2, 50ms	-0.26 (-0.5,0.0)
High-Freq. Amp	-0.58 (-0.8,-0.4)	F3, 10ms	-0.29 (-0.5,-0.1)
F1, 20ms	-0.48 (-0.7,-0.3)	F3, 50ms	-0.20 (-0.4,0.0)

Table 3.5: Normalized correlation between human transcriptions of the test data and measurement error of the automatic algorithm. 99% confidence limits are calculated using the Student's t distribution with 120 degrees of freedom.

μ_0 and μ_ϵ, their standard deviations σ_0 and σ_ϵ, and the normalized correlation coefficient

$$\rho = \frac{E[(x_0 - \mu_0)(\epsilon - \mu_\epsilon)]}{\sigma_0 \sigma_\epsilon} \tag{3.6}$$

The assumption of homoskedasticity is exactly the assumption that $\rho = 0$.

If we are given N tokens of the correct measurement x_0 and the measurement error ϵ, it is possible to estimate ρ by normalizing both x_0 and ϵ, and performing a one-tap, zero-mean linear regression:

$$\widehat{\rho} = \frac{1}{N} \sum_{i=0}^{N} \frac{(x_0(i) - \widehat{\mu}_0)(\epsilon(i) - \widehat{\mu}_\epsilon)]}{\widehat{\sigma}_0 \widehat{\sigma}_\epsilon} \tag{3.7}$$

where $\widehat{\mu}_0$, $\widehat{\sigma}_0$, $\widehat{\mu}_\epsilon$, and $\widehat{\sigma}_\epsilon$ are the sample mean and sample standard deviation of x_0 and ϵ, respectively.

The estimate $\widehat{\rho}$ is normally distributed with mean ρ, and the square of the residual $(\epsilon - \widehat{\rho} x_0)^2$ is distributed as a scaled χ^2 random variable (Johnson and Wichern, 1992). By appropriate calculations, we can derive $100(1 - \alpha)\%$ confidence limits on the value of the true correlation coefficient ρ:

$$|\rho - \widehat{\rho}| \leq t_{N-1}(\frac{\alpha}{2}) \sqrt{\frac{1 - (\widehat{\rho})^2}{N - 1}} \tag{3.8}$$

where $t_{N-1}(\alpha/2)$ is the $\alpha/2$ critical point of a Student's t statistic of order $N - 1$.

Table 3.5 shows the normalized correlation between several manually transcribed measurements and the corresponding measurement errors of the rule-based algorithm, estimated over the Error Modeling database. The hypothesis of homoskedasticity is rejected with 99% confidence for 6 out of the 10 measurements. Three measurements are much more strongly heteroskedastic than the others: the low-frequency and high-frequency burst amplitudes, and the measurement of F1 20ms after consonant release.

Figure 3-12 compares human and automatic transcriptions of the two burst amplitude measurements. According to the analysis in section 3.4.2, the standard deviation of the measurement error ϵ is 5-7dB for both measurements. If measurement error were uncorrelated with the correct measurement value, the lower plot in figure 3-12 should be a smeared version of the upper plot, with 5-7dB of variation added to the edges of the token cloud. Instead, the lower plot is slightly more compact than the upper plot. Low burst amplitude measurements have been increased, and, to a lesser extent, high burst amplitude measurements have been reduced. The error in this plot is clearly correlated with the measurements.

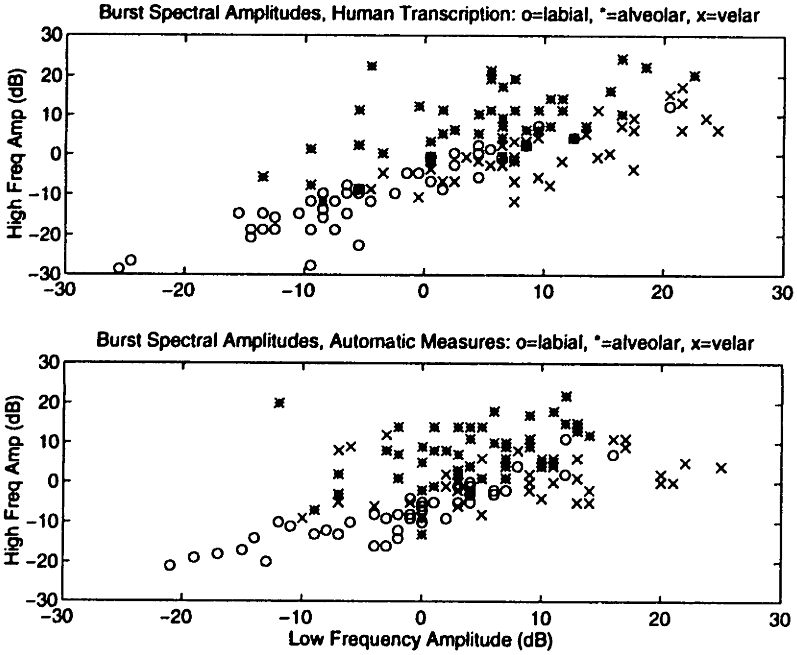


Figure 3-12: Burst spectral amplitudes, as measured by human transcribers, and by the rule-based measurement algorithm. Measurements by human transcribers have been shifted by a constant in order to make the plots comparable. The rule-based measurement algorithm has a significant tendency to reduce high-amplitude measurements, and increase low-amplitude measurements.

3.5 Discussion

This chapter develops and evaluates rule-based algorithms for the measurement of certain acoustic correlates of stop consonant place.

3.5.1 Algorithm training

The exact form of the algorithms developed here, and the values of various frequency and amplitude thresholds, were adjusted to imitate the performance of a human judge on a database consisting of 20 releases of each of the 9 stop and nasal consonants. The judge was given the identity and context of each stop, but did not use interactive spectral analysis to verify measurements, relying instead on spectral representations (time-averaged power spectrum, LPC roots) similar to those used by the automatic algorithms.

The performance of the algorithms on test data suggests that the algorithms were insufficiently trained. Compared to the measurements of judges on test data (generated interactively by judges using multiple spectral representations), the algorithm picks the wrong formant peak 10-20% of the time, and the wrong front cavity resonance 20% of the time.

Three reasons for the apparent fragility of the algorithms presented in this chapter can be suggested. Correction of these three problems is likely to lead to much more robust algorithms than those developed in this chapter.

First, the algorithms developed here use extremely simple statistical and speech production models, and as a direct result, the models are tuned quite closely to the distribution of the training data, and are therefore fragile to small variations. For example, the front cavity resonance measurement, which chooses the highest-frequency peak out of all those within 1dB of the spectral maximum, might be made more robust by the use of a more detailed model of the frequency dependence of the front cavity resonance. Similarly, the formant measurement would probably be more robust if it used a more global continuity constraint.

Second, the training data in this chapter are measured non-interactively, using the same spectral data used by the algorithm, while the test data are measured interactively, in order to encourage the best possible accuracy. As a result, it is possible that the training data measured by the human judge contain errors resulting from a poor spectral representation, some of which might have been corrected using interactive analysis. Apparently, limiting the human judge to the spectral representations of the algorithm is a mistake. Instead, the judge should be allowed to use as many different representations as he finds useful. If the algorithm is then unable to imitate the judge's measurements using a more limited spectral representation, then the input to the algorithm should also be supplemented, in order to obtain the most accurate measurements possible.

Finally, the training data set should be more closely matched to the test set. In this chapter, the training set consisted of only 20 releases of each consonant, in arbitrary and unbalanced phonetic contexts. As a result, unusual right contexts for each consonant were not considered in the algorithm design, and some of these contributed extra error to the test set. For example, /ky/ and /dw/ were not represented in the training data, and were found to contribute more than the average number of front cavity resonance errors in section 3.4.4.

Enthusiasts of speech production knowledge in the speech recognition community sometimes suggest that constraints based on speech production knowledge can be used as a substitute for additional training tokens. This may be true when the training database already consists of several thousand tokens, but when the training database is only a few hundred tokens, it is slightly misleading. In this chapter and the previous chapter, we have demonstrated that the *a priori* constraints imposed by speech production knowledge are often not very restrictive, but that the application of speech production knowledge as a continuous guiding principle in the development of measurement algorithms can be useful. Based on the development of measurements in this chapter, it might be suggested that speech production knowledge is most accurately characterized as knowledge which *guides the interpretation* of the available training tokens, helping us to better generalize the information available.

3.5.2 Aggregate error models

The introduction to this chapter suggests that automatic measurements are only useful if we have detailed models of the measurement error. This chapter has developed several detailed models of the aggregate statistical distribution of measurement errors. Aggregate models, however, have an important shortcoming: an aggregate error model makes the same error prediction for every utterance, with no regard for the huge range of spectral differences between utterances. This section discusses the extent to which aggregate error models might be useful in the applications of speech sound classification and inference of articulatory correlates. The discussion here concludes by proposing a tighter link between the measurement algorithm and the error model; this link will be developed further in the next chapter.

Classification

Most parametric classifiers, including linear discriminant and mixture Gaussian classifiers, use a feedforward classification algorithm: all relevant measurements are presented to the classifier as a single measurement vector, and a classification score is computed as an algorithm-specific function of the given measurements. There is no way to incorporate an aggregate model of measurement error directly into a parametric classifier, since everything the classifier knows about variability in the measurements is represented in the (algorithm-specific) classifier weights.

It is possible, however, to combine parametric classifiers and error models under the supervision of a higher-level program. Carver and Lesser (1992), for example, have developed a non-speech sound recognizer using an expert system. In their system, signal processing "knowledge sources" contribute knowledge about the error inherent in different signal representations, and the expert system uses this knowledge to decide which representations to use for a given classification task. Johnson (1994) proposed selecting the best measurements for a given classification task on the basis of phonetic context, using a table lookup scheme, and Chun (1996) developed a system which selects measurements using a decision tree; both of these systems could be modified to incorporate knowledge about measurement error.

In all of these systems, spectral representations are not measured until they are requested by the supervisor program. If measurement errors are compactly distributed (for example, if errors are Gaussian), and if the error standard deviation is small compared to the range of measurements characterizing a phoneme, this kind of on-demand measurement allows significant computational reduction. An expert system classifier using Gaussian error models, for example, might proceed as follows. First, an acoustic measurement is requested, and the value of the measurement is classified. If the value of the measurement is sufficiently close to a category boundary that it might have been pushed across by measurement error, another measurement is requested. The process continues, with more measurements performed and classified, until the possibility of a misclassification caused by measurement error reaches an acceptably low level. In this way, time-consuming acoustic measurements are only used in the classification of tokens very close to a category boundary.

If error is not compactly distributed — that is, if there are many outliers — aggregate error modeling is not as useful. For example, in the concentric mixture Gaussian models of section 3.4.3, any frequency measurement has a 10-20% chance of being an outlier. Outliers are often large enough to cause classification errors; even a measurement which is far from any category boundary can be shifted across the boundary by a sufficiently large measurement error. Thus, an expert system using a concentric mixture Gaussian error model is forced to assume that any measurement, regardless of how close it is to a category boundary, has a 10-20% chance of being misclassified because of measurement error.

Measurement of probability distributions

Section 2.4.2 suggests that it may be possible to infer the statistical distribution of an articulatory parameter from the measured distribution of its acoustic correlates. If a measured acoustic distribution contains errors, however, the inferred articulatory distribution will contain similar errors. The power of automatic acoustic measurements to predict articulatory measurements therefore depends critically on the reliability of the measured acoustic distribution.

Suppose x_0 is the true value of an acoustic correlate measurement, and $x = x_0 + e$

is a possibly erroneous automatic measurement of the same acoustic correlate. If e is independent of x_0, [1] the probability density $p_x(\xi)$ is the convolution of the densities $p_{x_0}(\xi)$ and $p_e(\xi)$, where convolution is denoted by $*$, and ξ is a dummy variable:

$$p_x(\xi) = p_{x_0}(\xi) * p_e(\xi) \tag{3.9}$$

If the error distribution $p_e(\xi)$ is compact (e.g. Gaussian), the convolution in equation 3.9 is just a smoothing operation, and the distribution $p_x(\xi)$ of the automatic measurement is just a smoothed version of the distribution $p_{x_0}(\xi)$ of the correct measurement. Suppose y_0 is an articulatory parameter related to x_0 by a differentiable function $x_0 = f(y_0)$. Then the probability distribution p_{y_0} (Zwillinger, 1996) is

$$p_{y_0}(\xi) = p_{x_0}(f(\xi))\frac{df(\xi)}{d\xi} \tag{3.10}$$

If the smoothed distribution p_x is used in equation 3.10 in place of p_{x_0}, the estimated articulatory distribution $p_y(\xi)$ is just a nonlinearly smoothed version of the true distribution $p_{y_0}(\xi)$.

If the error distribution $p_e(\xi)$ contains outliers (e.g. a concentric mixture Gaussian), then equation 3.9 no longer represents a local smoothing operation. For example, if the error distribution is a concentric mixture Gaussian composed of a compact Gaussian distribution $p_c(\xi)$, chosen with probability $1 - P$, and an outlier distribution $p_o(\xi)$, chosen with probability P, then the distribution $p_x(\xi)$ is

$$p_x(\xi) = (1 - P)(p_{x_0}(\xi) * p_c(\xi)) + P(p_{x_0}(\xi) * p_o(\xi)) \tag{3.11}$$

The first term on the right in equation 3.11 represents a locally smoothed version of p_{x_0}, similar to the distributions considered above. The second term is also a smoothed version of p_{x_0}, but the smoothing is over such a wide range that most features of p_{x_0} are lost, and the result looks something like a constant noise floor added to the total distribution $p_x(\xi)$.

Thus we find that, if errors are compactly distributed, the distributions of acoustic and articulatory parameters estimated using erroneous measurements are just smoothed copies of the actual distributions. An error distribution containing outliers, on the other hand, adds a sort of "noise floor" over a broad parameter range to both the acoustic and articulatory parameter distributions, which may mask details of the true parameter distribution.

Aggregate versus individual error distributions

Up to this point, this section has discussed the combination of automatic measurements and aggregate error models for use in the applications of phonetic classification and articulatory parameter estimation. Compact (e.g. Gaussian) error distributions have been shown to produce a locally smoothed measurement distribution, corresponding to a local band of possibly misclassified tokens near the category boundary of a phonetic classifier. Error distributions with outliers, on the other hand, have been shown to add a roughly constant noise floor to the measurement distribution, and to add a corresponding roughly constant probability of misclassification to all tokens considered by a phonetic classifier. Since all of the frequency measurements considered in this chapter have significant outliers, it must be

[1] A heteroskedastic error term requires a few extra calculations, which are not reproduced here.

concluded that classification and articulatory inference based on these measurements would suffer from a relatively constant noise floor, which would presumably limit the usefulness of these measurements in either application.

All of this discussion assumes that the only thing known about a given measurement error is that it is randomly drawn from a particular distribution. To anyone who has ever attempted manual phonetic transcription, however, this assumption is absurd. When a human attempts to measure formant frequencies or a front cavity resonance, most of the measurements he makes are easy, because the formant or front cavity resonance peak is clearly visible in the expected frequency band. Most of the errors made by a human transcriber come from a small number of atypical spectra, in which there are either multiple candidate formant peaks, or no candidate formant peaks. Furthermore, a human transcriber can usually tell which tokens are difficult to measure, and provide a rough estimate of the probability that any given formant or front cavity resonance measurement contains a gross error.

Token-by-token estimates of the probability of error would be tremendously useful in an expert system speech classifier, and possibly in the inference of articulatory measurements. If most formants can be measured with little probability of gross error, an expert system classifier can focus resources (in the form of additional measurements) on the tokens which are most likely to contain errors. Likewise, an algorithm estimating the histogram of an articulatory parameter can weight the acoustic input tokens by their probability of correctness, so that a single gross measurement error has less influence on the estimated parameter distribution.

Although automatic measurement algorithms are more susceptible to error than human transcribers, most gross automatic measurement errors seem to come from the same kind of ambiguous spectra that produce human errors. Furthermore, estimation of the probability of error for each measurement seems to require spectral information which is similar to the kind of information used to make the measurement in the first place: for example, knowledge of the frequencies and amplitudes of convex peaks in a burst spectrum is important for both measuring the front cavity resonance, and for estimating its probability of error. Processing this information in order to produce an estimate of error, however, requires a degree of self-supervision which the algorithms developed in this chapter do not possess.

The next chapter develops a formant tracker which uses hidden Markov modeling (HMM) technology to estimate the spectral uncertainty inherent in each formant measurement.

The introduction of HMM formant tracking into an otherwise knowledge-based thesis has sparked philosophical complaints from a few reviewers. Historically, the field of speech recognition has occasionally been divided into the camps of knowledge-based and HMM-based recognition, where the two camps can be caricatured by saying that engineers using HMMs want the biggest possible training database, while engineers using knowledge-based systems think that production knowledge obviates the need for extra training data. The author of this thesis strongly believes that the historical division of speech recognition into these two camps is an anachronism, which only serves to keep engineers from seeking the best solutions to speech recognition problems.

Section 3.5.1 proposes that speech production knowledge is best characterized not as a substitute for training data, but rather as an aid in the interpretation of training data. Continuing this line of reasoning, chapter 4 uses HMMs, trained on data, to develop a formant tracker capable of estimating its own measurement uncertainty. This trained HMM formant tracker is not a substitute for knowledge-based formant trackers; it *is* a knowledge-based formant tracker. Speech production knowledge contributes to the choice of formant frequencies as production states, to the choice of a bell-shaped formant transition probability, and to the design of the spectral input parameters. In fact, it might be argued that the HMM formant tracker in the next chapter incorporates more speech production knowledge than the rule-based tracker developed in this chapter, because the HMM formant tracker uses a global rather than a local continuity constraint.

Chapter 4

Stochastic Formant Modeling

This chapter demonstrates a hidden Markov formant tracker which allows us to model the uncertainty in each formant measure, individually, as the formant is measured. Several examples of formant uncertainty are presented, caused, for example, by multiple peaks in the expected formant range, or by the disappearance of a formant peak. Methods for evaluating the uncertainty model are discussed.

4.1 Phoneme-Dependent HMM Formant Tracking

A strictly bottom-up measurement algorithm, like those described in the previous chapter, can only make use of context information which is already known. In contrast, a Bayesian approach, such as a hidden Markov model (Kopec, 1986), can be used to explicitly test different hypotheses about the unknown information. If the goal is an estimate of the formant frequencies, we can identify the formant frequencies which are most likely given the phonetic transcription. If the goal is recognition of consonant place, we can measure the formants that best fit the model of each consonant place, and then pick out the model which best explains the data.

This chapter describes a phoneme-dependent hidden Markov model formant tracking algorithm. In the classification experiments of chapter 5, this model includes a two-level search space, in which the formant measurements and the phoneme sequence are identified simultaneously, as shown in figure 4-1. In contrast, all of the experiments in this chapter assume that the identity and release time of the consonant are known perfectly, and formant tracking will make use of that knowledge.

4.1.1 A phoneme-dependent transition model

In an HMM formant tracker, the formant frequencies are treated as hidden states of the production mechanism, rather than explicitly observable features of the spectrum (Kopec, 1986). In the model developed here, as in Kopec's model, the frequency scale is divided into discrete formant states, and formant continuity over time is regulated by a bell-shaped transition probability. Unlike Kopec's model, the model developed here does not allow formants to "disappear" occasionally. Instead, formants are required to exist at some real frequency at all times, and occasional disappearances of the corresponding spectral peak are handled by the output model. This change has little practical consequence, but follows naturally from the philosophical view that a formant, as a natural frequency of the physical

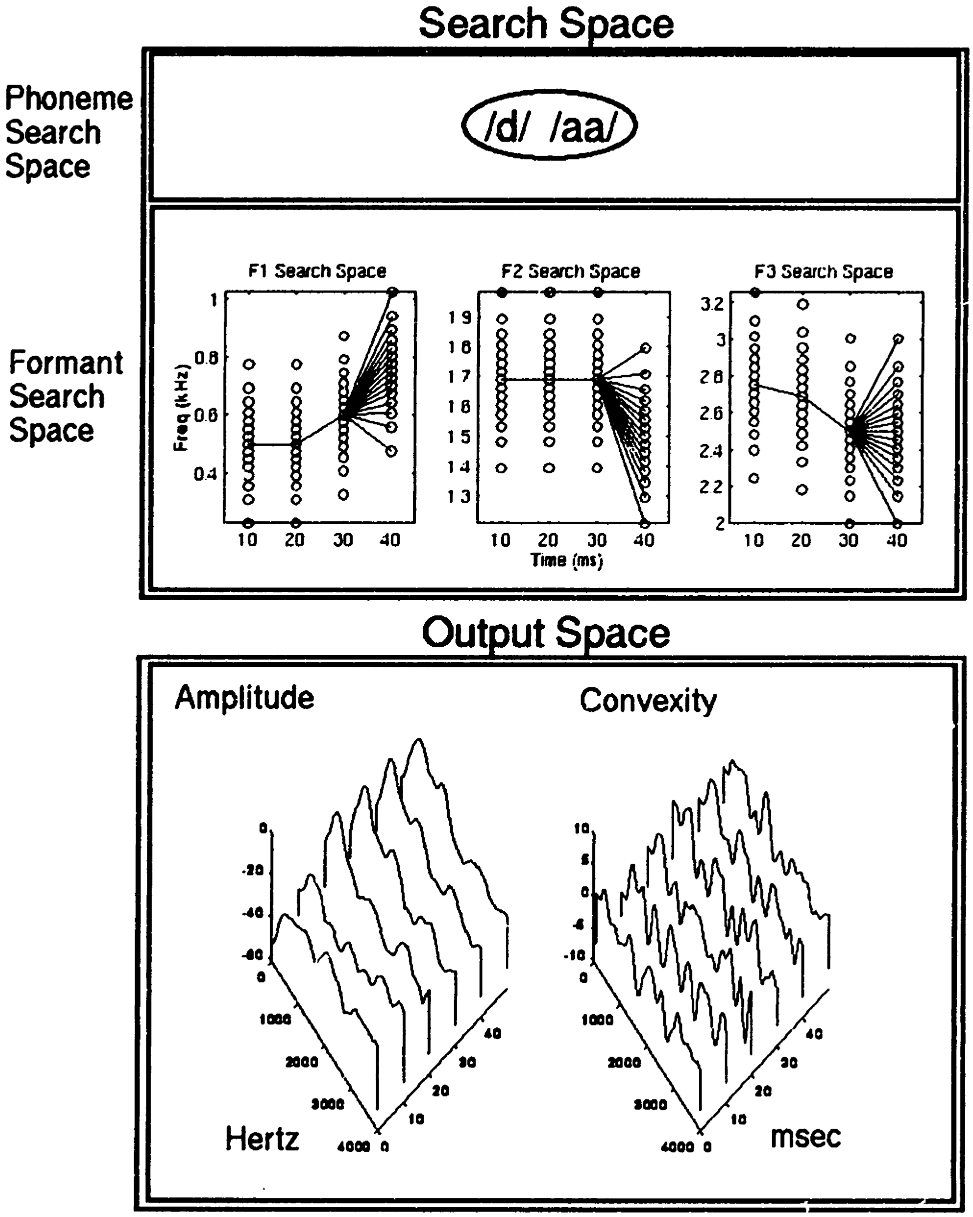


Figure 4-1: Schematic representation of a phoneme-dependent hidden Markov model formant tracker. Time of the landmark is assumed given. The search space consists of a phoneme level, and a formant frequency level. Formant tracks are semi-continuous in frequency, and discrete in time, with a formant frequency space searched once every 10ms. Spacing of the circles in the schematized formant frequency search space, above, represents the height of the probability density curve as a function of frequency at each time. The output space consists of models of spectral amplitude and convexity as a function of formant frequency, phoneme identity, and time. Spectral amplitude and convexity are computed using appropriate transformations of the periodic time-averaged power spectrum, as described in the text.

vocal tract, must have a real value at all times, whether or not it is represented in the acoustic spectrum.

The most important difference between the model proposed here and Kopec's model is the introduction of phonemic information. In the model proposed here, formant transition probabilities are a function of time (relative to the consonant release), and of the underlying phoneme sequence. This modification is appropriate in two applications: formant tracking with a known phonemic transcription, and phoneme classification with a known release time. Formant tracking with a known transcription is considered in this chapter. Phoneme classification, using the two-level search space shown in figure 4-1, is considered in chapter 5.

A formant tracker with transition probabilities conditioned on both time and phoneme label is much more complex than a phoneme-independent tracker. In order to be trainable, the basic transition model must be as simple as possible. In the simplest appropriate model, each formant is modeled as an independent continuous-state discrete-time Gauss-Markov process. The parameters of this process are the mean formant frequency as a function of time $\mu_f(t,S)$, the formant variance $\sigma_f^2(t,S)$, and the inter-frame normalized correlation coefficient $\rho_f(t,S)$, all of which may be functions of time t and of the underlying phoneme string S:

$$p(F_n(t)|F_n(t-1)) = N\left(\frac{F_n(t)-\mu_f(t,S)}{\sigma_f(t,S)\sqrt{1-\rho_f^2(t,S)}} - \rho_f(t,S)\frac{F_n(t-1)-\mu_f(t-1,S)}{\sigma_f(t-1,S)\sqrt{1-\rho_f^2(t,S)}}\right) \tag{4.1}$$

where $N(x)$ is the unit normal distribution.

4.1.2 A DFT-based output model

In recognition using a hidden Markov model, state variables are identified by inverting a hypothetical "production model" which specifies the contingent probability of a spectral sequence given the state. In the model proposed here, the state variables are formant frequencies; phoneme identity and release time are implicit state variables, which are assumed to be known. The observation sequence consists of the amplitude $A(f,t)$ and convexity $C(f,t)$ of a time-averaged power spectrum. This section defines the amplitude and convexity, and describes a model of the output probabilities.

To reduce complexity, the output model used here adopts two important simplifications. First, spectra are calculated using a linear frequency axis, because a linear axis simplifies modeling and analysis of formant motion, and despite the fact that there is considerable evidence supporting the use of mel-scale or bilinear frequency warping in speech recognition (e.g. Davis and Mermelstein, 1980). Second, the influence of a formant is assumed to be strictly local: the influence of a formant $F_n(t)$ on the output spectrum $\mathbf{O}(t)$ is limited to changes in the local amplitude $A(F_n,t)$ and convexity $C(F_n,t)$, so that

$$p(F_n(t),\mathbf{O}(t)) = p(F_n(t),A_n(F_n,t),C_n(F_n,t)) \tag{4.2}$$

The spectral amplitude at frequency f is defined as the ratio of the time-averaged power spectral amplitude $P_x(f,t)$ to the total DFT energy $E(t) = \int_0^{f_s} P_x(f,t)df$. Since ratio variables are not well modeled by a Gaussian distribution, the ratio $P_x(f,t)/E(t)$ is

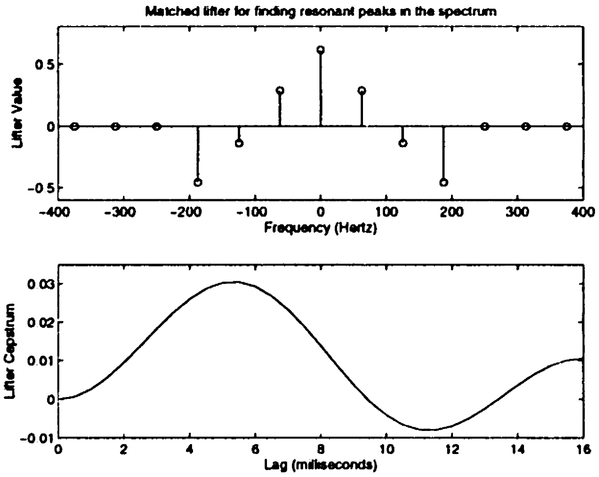


Figure 4-2: Lifter used to estimate spectral convexity. The lifter is a seven-sample match to the spectrum of a 125Hz-bandwidth spectral resonance, normalized to zero mean and unit energy. The resulting lifter has a bandpass cepstrum, amplifying spectral features which have half-periods between roughly 67 and 167 Hertz.

transformed using the logistic transform,

$$A(f,t) \equiv 10\log_{10}\left(\frac{P_x(f,t)}{E(t) - P_x(f,t)}\right) \tag{4.3}$$

where the normal formula for the logistic transform (e.g. Johnson and Wichern, 1992) has been scaled to return an output in decibels.

The convexity measurement $C(f,t)$, computed using the lifter shown in figure 4-2, is an approximation of the second derivative of the log spectrum as a function of frequency. Even when two formants merge (section 2.2.5), it is often possible to find the hidden formant by convolving the spectrum with a short FIR lifter designed to pick out the hidden formant resonance curve. The lifter used in this chapter, and its cepstrum, are shown in figure 4-2. The FIR lifter, shown in the upper panel, was designed by using a 400Hz rectangular window to truncate the spectrum of a 125Hz-bandwidth complex pole pair; the resulting seven-sample lifter was normalized to have zero mean and unit energy. The lifter has a band-pass cepstrum, shown in the lower panel of figure 4-2, which amplifies spectral features with half-periods between roughly 67 and 167 Hertz (roughly the bandwidth of a formant, although high frequency formants may have larger bandwidths).

For simplicity, we would like to model the distribution of convexity measurements $C(f,t)$ using a Gaussian distribution. A Gaussian model, however, has the undesirable characteristic of symmetry: convexity measurements larger than the mean are as unlikely as measurements smaller than the mean. When human judges measure formant frequencies, they usually locate a formant frequency near a local maximum of $C(f,t)$, implying that the convexity likelihood function should be non-decreasing. The experiments described in this chapter model convexity using a "half-Gaussian" distribution, in which all convexity values above a trained mean parameter are considered equally likely. The half-Gaussian is not really a probability density, since it does not integrate to unity; in chapter 5, when normalized probabilities become important, a different model will be introduced.

To summarize, the output model used in this chapter is local, in the sense that a formant at frequency F_n is assumed to affect only the amplitude $A(F_n, t)$ and convexity $C(F_n, t)$ at the same frequency. The amplitude, after normalization and a logistic transform, is modeled using a Gaussian distribution; the convexity is modeled using a half-Gaussian. The mean $\mu(t, S, n)$ and standard deviation $\sigma(t, S, n)$ of both of these distributions are allowed to depend on time t (relative to the landmark), the phoneme string S, and the formant number n (but not the formant frequency):

$$p(\mathbf{O}(t)|F_n(t)) = N\left(\frac{A(F_n,t) - \mu_A(t,S,n)}{\sigma_A(t,S,n)}\right) N\left(\min\left(\frac{C(F_n,t) - \mu_C(t,S,n)}{\sigma_C(t,S,n)}, 0\right)\right) \quad (4.4)$$

4.2 Estimating Parameters from a Reference Transcription

The simplified formant model developed above consists of a Gauss-Markov continuity model, with three trainable parameters, and two Gaussian output distributions, with two trainable parameters each. A consonant release model consisting of six spectral frames, matched to three formant models, requires training or interpolation of $6 \times 3 \times 7 = 72$ parameters.

The parameters of 18 consonant release models (9 consonants × 2 speaker genders) were trained based on human transcriptions of a small training database. The SFM Train database (described in appendix A.3) consists of 4 tokens of each of 9 consonants, released into the vowels /aa/ and /ah/. Data from these two vowels were pooled for training, on the assumption that their formant frequencies within 50ms of consonant release should be similar.

In order to make the trained model parameters as reliable as possible, given the limited training data, the data were grouped according to relevant features before training. The grouping for each trainable parameter was slightly different, as described in the following two sections.

4.2.1 Means as a function of time

According to the speech production theory sketched in chapter 2, the vocal tract formant frequencies at release of a consonant should depend little on the manner of the consonant. The manner of the consonant may affect the excitation of a formant, and therefore its amplitude and convexity, but the underlying vocal tract resonant frequencies should depend only on consonant place. Conversely, gender of the speaker is unlikely to affect the amplitude and spectral convexity of a formant, but it certainly affects the formant frequency.

Mean formant frequencies were trained by grouping together tokens with the same place of articulation and the same speaker gender: a total of six training tokens for each of six independent training cells. Mean amplitudes and convexities were trained separately for each of the nine consonants, regardless of speaker gender, with a total of four training tokens for each consonant.

For training sets of this size, the possibility of incorrect training is significant. Assuming that the training tokens for each consonant are drawn from an underlying Gaussian distribution, the standard error of the frequency parameter estimates is $1/\sqrt{6}$ times the standard deviation, and the standard error of amplitude and convexity parameter estimates is 1/2 the standard deviation.

Fortunately, the number of models trained in this section is sufficiently small to allow visual confirmation. The mean formant frequencies, as trained on the 36 training tokens, are shown in figure 4-3.

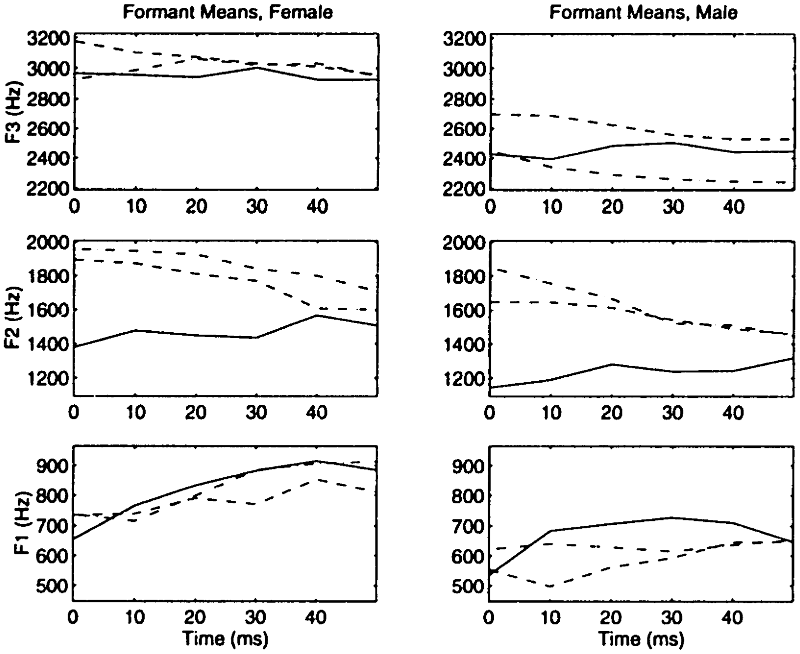


Figure 4-3: Mean formant frequencies for three places of articulation, for each gender, trained on 36 tokens. Solid=labial, dashed=alveolar, dot-dash=velar.

Random variation of the formant tracks in figure 4-3 appears to be less than about 100Hz in most cases, and most large-scale features are as they should be. Perhaps the only unusual feature of both the male and female distributions is the large gap between the velar F2 onset and the velar F3 onset. Visual examination of the training tokens indicates that most velar tokens had relatively low F3 onsets, but that a few tokens with unusually high onsets pulled up the means. Perhaps, with more training data, a mixture Gaussian model would be a more appropriate representation of the velar F3 onset than a simple Gaussian.

Figure 4-4 shows amplitude and convexity means, averaged across both voiced and unvoiced stops. There is quite a bit of random variation in this plot. As expected, the onset of A2 is highest for velar stops, because of the velar front cavity resonance at the frequency of F2. Also as expected, the convexity C1 rises as a function of time for all places of articulation. An unexpected feature of the amplitude tracks is the high A1 onset of both alveolar and velar stops, which can be attributed to strong subglottal resonances and/or voicing information in the vicinity of F1 for several of the stops. Place-dependent differences in convexity are apparently not significant, and are generally much less than the standard deviation.

4.2.2 Variance and covariance

It is impossible to train reliable variance estimates on the basis of four to six training tokens; assuming an underlying Gaussian distribution, the variance estimate is a scaled χ^2 random variable with a standard error equal to $1/\sqrt{2}$ to $1/\sqrt{3}$ times its expected value.

Estimated variance parameters were therefore not allowed to vary with time. The variance as a function of time of each parameter was estimated using the groupings previously described, and then these variance tracks were averaged over the six time frames in each token. The standard deviations of the spectral amplitude calculated in this way were be-

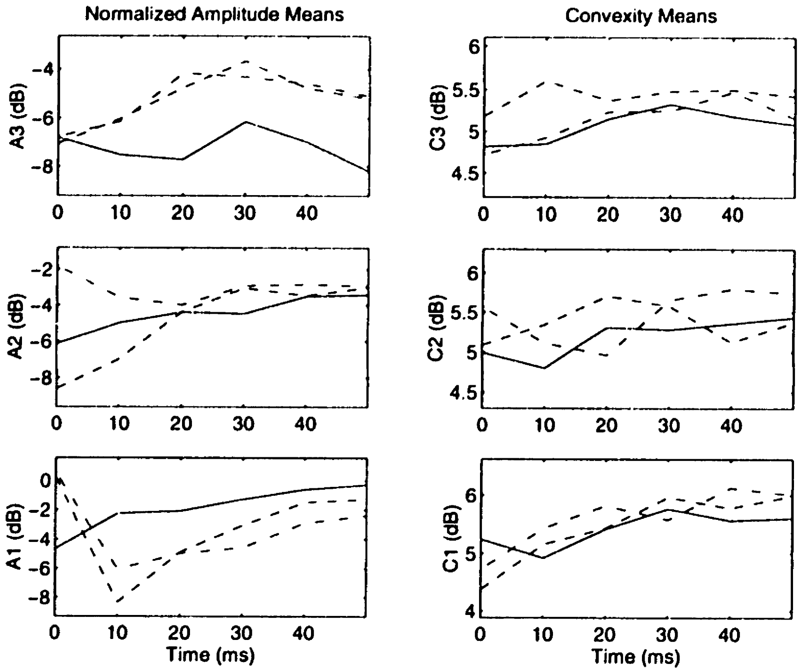


Figure 4-4: Mean amplitude and convexity, averaged across all stop tokens, for three places of articulation. Solid=labial, dashed=alveolar, dot-dash=velar.

tween 2.4 and 6.5dB, and were usually larger for unvoiced stops than for either voiced stops or nasals. The standard deviations of the spectral convexity were between 0.9 and 1.8, with large and small values apparently distributed at random among the phonemes and formants.

After being averaged over time, the variances of the formant frequencies were, finally, averaged over both place and gender, in order to produce global formant variance estimates. The model standard deviations of F1, F2, and F3 are, respectively, 184, 221, and 303 Hertz.

The temporal correlation coefficient, ρ_f, was set to a single constant for all formant frequencies. The HMM formant tracking algorithm was tested on the 36 training tokens with several values of this parameter, and the value $\rho_f = 0.88$ was found to produce the best fit between the *a posteriori* formant distributions, described below, and the transcribed formant frequencies.

4.3 Real-Time Estimation of Measurement Uncertainty

The search space of a hidden Markov model is usually explored using one of two standard dynamic programming algorithms (Rabiner, 1993): the Viterbi algorithm, and the forward-backward algorithm. In an HMM formant tracker, the Viterbi algorithm might be used to identify the single most likely set of formant tracks for a given utterance. The forward-backward algorithm, on the other hand, can be used to calculate the *a posteriori* probability of every possible formant frequency in each time frame, given an utterance and a trained model.

The *a posteriori* state probabilities produced by the forward-backward algorithm can be viewed as frame-by-frame models of the measurement uncertainty of the algorithm. If a formant has a clearly defined spectral peak at the expected frequency, the *a posteriori* probability distribution is well localized. Conversely, if there are no formant-like peaks in the expected frequency range, or if there are several, the location of the formant is uncertain, and the *a posteriori* probability is more diffuse (see the examples in section 4.3.2, below).

4.3.1 Finding a posteriori formant distributions

The *a posteriori* probability of a formant frequency $F_n(t)$, given a vector of model parameters λ and a finite sequence of observed spectra $\mathbf{O}(0)\ldots\mathbf{O}(T)$, is defined to be

$$p(F_n(t)=f|\lambda,\mathbf{O}(0),\ldots,\mathbf{O}(T))=\frac{p(f,\mathbf{O}(0),\ldots,\mathbf{O}(T)|\lambda,t)}{\sum_g p(g,\mathbf{O}(0),\ldots,\mathbf{O}(T)|\lambda,t)} \tag{4.5}$$

In the forward-backward algorithm, the joint probabilities $p(f,\mathbf{O})$ on the right hand side of equation 4.5 are divided into two factors, and these factors are computed recursively using an efficient algorithm. The two factors are called the forward probability $\alpha_n(f,t)$ and the backward probability $\beta_n(f,t)$:

$$p(F_n(t),\mathbf{O}(0),\ldots,\mathbf{O}(T)|\lambda)=\alpha_n(F_n,t)\beta_n(F_n,t) \tag{4.6}$$

$$\begin{aligned}\alpha_n(F_n,t)&\equiv p(F_n(t),\mathbf{O}(0),\ldots,\mathbf{O}(t)|\lambda)\\ \beta_n(F_n,t)&\equiv p(\mathbf{O}(t+1),\ldots,\mathbf{O}(T)|\lambda,F_n(t))\end{aligned}$$

The forward and backward probabilities can be computed recursively according to the forward-backward algorithm. If we use a strictly local output model, as described in equation 4.2, the forward-backward recursion formulas are:

$$\alpha_n(F_n,t)=p(A_n|F_n,t)p(C_n|F_n,t)\sum_{F_n(t-1)}p(F_n(t)|F_n(t-1))\alpha_n(F_n,t-1) \tag{4.7}$$

$$\beta_n(F_n,t)=\sum_{F_n(t+1)}p(F_n(t+1)|F_n(t))p(A_n|F_n,t+1)p(C_n|F_n,t+1)\beta_n(F_n,t+1)$$

4.3.2 Examples

Figure 4-5 shows the spectrum, spectral convexity, and posterior formant probabilities at 10ms and 50ms after the /b/ release in the word "Barb" spoken by a female speaker. *A posteriori* formant probabilities are calculated with the forward-backward algorithm, using an observation sequence consisting of six spectra between the release and 50ms after release.

Notice that when there is a clear, narrow formant peak in the spectrum, the *a posteriori* distribution of the corresponding formant is tightly confined. As F1 moves up from the first spectrum to the second, it obscures the location of the F2 peak, and as a consequence, the *a posteriori* distribution for F2 at time 50ms is more diffuse.

A different kind of uncertainty is visible in the F3 region at time 10ms. Here, there are two peaks close together (possibly F3 and F4). The forward-backward algorithm tests both peaks, to see how well they fit into the global F3 track being developed throughout the syllable, and assigns them *a posteriori* probabilities on this basis.

Figure 4-6 shows the spectrum, spectral convexity, and *a posteriori* formant probabilities 10ms and 50ms after release of the /d/ in "dark," spoken by a different female speaker.

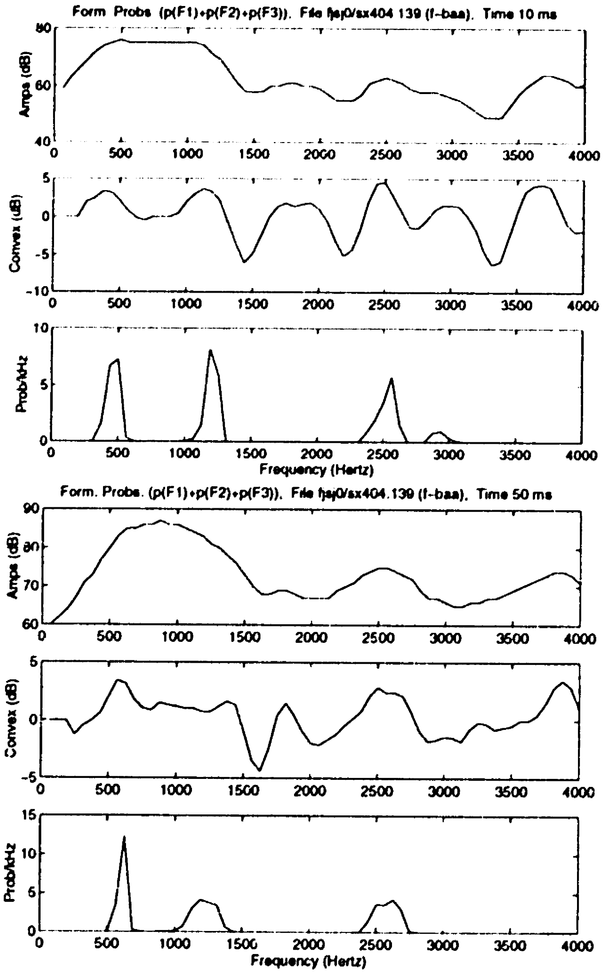


Figure 4-5: Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 10ms and 50ms after release of the first /b/ in "Barb." *A posteriori* probability distributions of all three formants are shown on the same plot, because there is no significant overlap between the distributions.

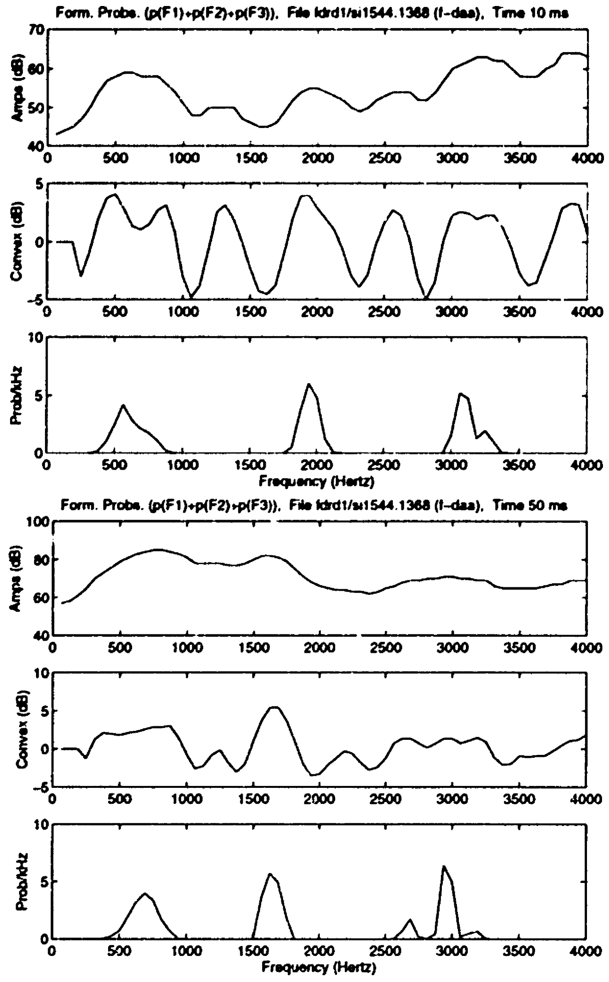


Figure 4-6: Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 10ms and 50ms after release of the /d/ in "dark." *A posteriori* probability distributions of all three formants are shown on the same plot, because there is no significant overlap between the distributions.

Notice, again, that the *a posteriori* probability is more diffuse when there is more than one possible formant location.

The *a posteriori* distribution of F3 in this figure demonstrates two problems with the formant tracker. First, at 10ms after release, the formant tracker seems to split the probability density of F3 between two distinct frequencies, at about 3050 and 3250Hz. These two peaks in the *a posteriori* distribution correspond to distinct peaks in the convexity measure, but the DFT spectrum itself seems to have only one peak, centered at about 3200 Hertz. Apparently, the single DFT peak is too broad to be identified as a single peak by the convexity lifter, indicating that possibly the passband of the convexity lifter is too narrow.

The second tracking difficulty in figure 4-6 occurs at 50ms after release. F3 is at about 2700 Hertz in this spectrum, and continues to fall into the /r/ as the syllable progresses. The formant tracker, however, only has information about spectra between the release and 50ms after release. Since the formant tracker has no way of knowing that F3 will continue to fall, it also has no reliable way of distinguishing between the peaks at 2700, 3000, and 3200 Hertz. The peaks at 3000 and 3200 Hertz are both continuous with apparent peaks in the release, but 3200 Hertz is an unusually high frequency for F3, so the algorithm assigns most of the *a posteriori* F3 probability to the peak at 3000 Hertz.

4.4 Evaluating the Error Models

The *a posteriori* formant uncertainty models described above were tested using the 36 tokens containing /aa/ and /ah/ in the Error Modeling database, described in appendix A.2.

4.4.1 Cumulative-probability representation of reference measurements

Since the uncertainty model varies from token to token, its ability to predict true formant location must be evaluated on a per-token basis. Since the uncertainty model is not a parametrized function of frequency, the method of evaluation should be non-parametric. This section describes a non-parametric, token-by-token extension of the standard Q-Q plot (Johnson and Wichern, 1992) which can be used to evaluate a real-time uncertainty model.

To the extent that a measurement uncertainty model is correct, the model should predict the distribution of possible frequencies of the true formant. Thus, regardless of the shape of an *a posteriori* distribution, the true formant should fall in the bottom N% of the distribution roughly N% of the time.

For a given observation sequence $\mathbf{O}(0), \ldots, \mathbf{O}(T)$, the degree of correspondence between the uncertainty models $p(F_n(t) = f|\mathbf{O}(0), \ldots)$ and a reference transcription containing the true formants $F_n(t)$ can be characterized by a cumulative probability:

$$P_{f\leq}(F_n(t)|\mathbf{O}(0), \ldots) = \int_0^{F_n(t)} p(F_n(t) = f|\mathbf{O}(0), \ldots)df \qquad (4.8)$$

Equation 4.8 contains a continuous integration, but the uncertainty model generated by an HMM formant tracker is discrete. The experiments reported here approximate equation 4.8 by integrating a trapezoidal interpolation of the discrete HMM uncertainty model.

The conversion from formant measurement to cumulative probability is shown in figure 4-7. In the figure, the reference formant measurement transcribed by a human judge is F2=1616 Hertz. The *a posteriori* probability distribution generated by the forward-

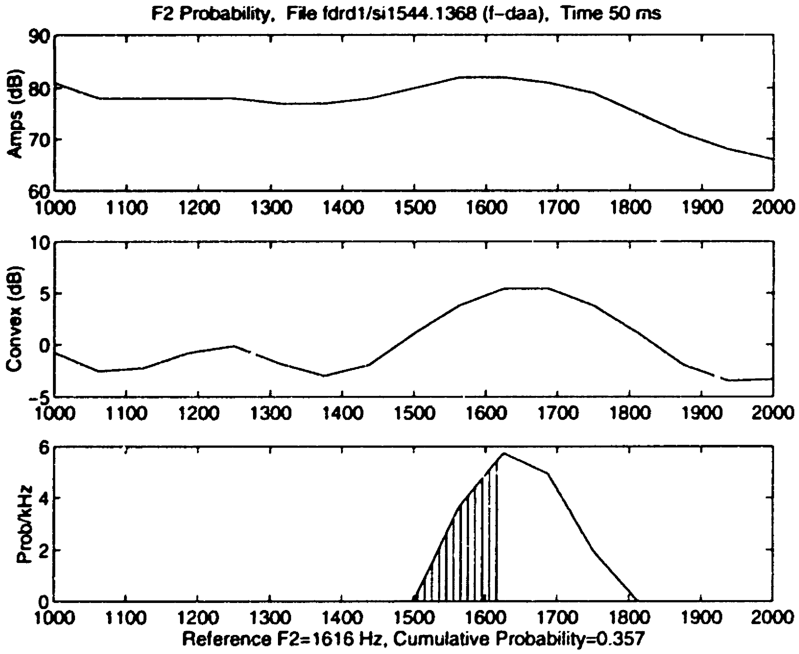


Figure 4-7: Comparison of a known formant frequency to the uncertainty model generated by an HMM formant tracker. F2 is located at 1616 Hertz, according to the measurements of a human transcriber. In the uncertainty model generated by the HMM, the probability of an F2 measurement less than or equal to 1616 Hertz is 35.7%; thus we can say that the cumulative probability representation of the human transcription is $P = 0.357$.

backward algorithm predicts that there is a 35.7% probability that F2 is less than or equal to 1616 Hertz, so the cumulative probability representation of the human transcription is $P_{f\leq}(1616|\mathbf{O}(0),\ldots) = 0.357$.

4.4.2 Evaluation using a binomial distribution

To the extent that the uncertainty models are correct, the cumulative probability measurements $P_{f\leq}(F_n(t)|\mathbf{O}(0),\ldots)$ should be uniformly distributed between 0 and 1.

The uniformity of distribution of a large number of samples can be tested by sorting them into the bins of a histogram. If the bins are uniformly spaced, there should be an equal number of tokens in each bin. Specifically, if N tokens are uniformly distributed into M bins between zero and one, the number of tokens n_j in bin number j is a binomial random variable, with mean N/M and variance $N(M-1)/M^2$.

Given M bins and N tokens, the last bin count n_M is a linear function of the other variables:

$$n_M = N - \sum_{i=1}^{M-1} n_i \tag{4.9}$$

If we arrange the bin counts into a vector, $\mathbf{n}_M = [n_1, \ldots, n_M]^T$, the covariance matrix $C_M = E[\mathbf{n}_M \mathbf{n}_M^T]$ is singular. The largest non-singular submatrix of C_M is the $M-1$-dimensional submatrix C_{M-1}, formed by eliminating the last row and column of C_M:

$$\mathbf{C}_{M-1} = E[\mathbf{n}_{M-1}\mathbf{n}_{M-1}^T] = \frac{N}{M}\left(\mathbf{I} - \frac{\mathbf{U}}{M}\right) \tag{4.10}$$

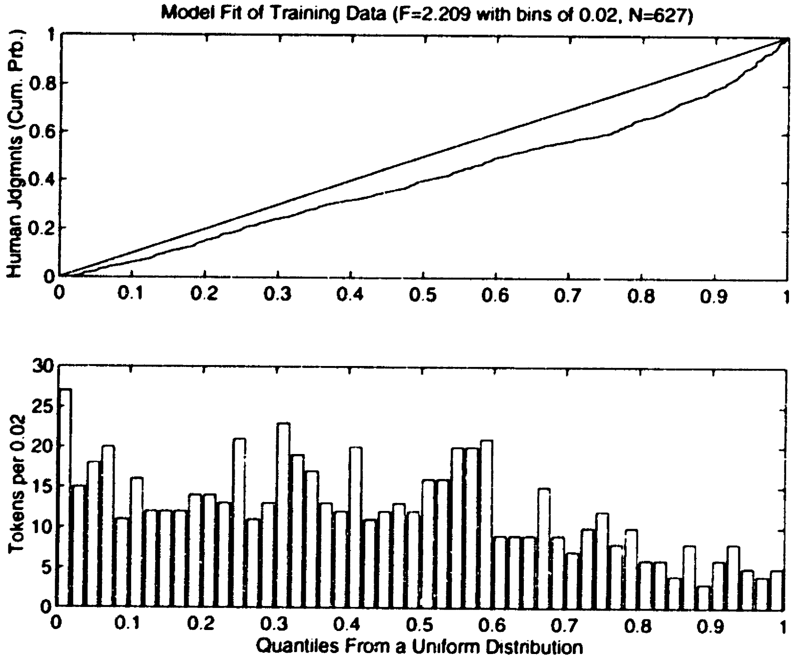


Figure 4-8: Cumulative probability representations of the human formant measurements used to train the formant tracker (F1, F2, and F3). Deviation from a uniform distribution indicates that the formant tracker is not adequately representing variation in the training data.

where **I** is the identity matrix, and **U** is an $M-1 \times M-1$ matrix of ones.

The binomial distribution approaches the Gaussian distribution rapidly as N increases. To the extent that the Gaussian approximation holds, we can evaluate the fit between the transcribed formant measurements and the *a posteriori* model distributions by using standard statistical tools to compare $\mathbf{n}_{M-1}$ to its expected value, $E[\mathbf{n}] = (N/M, \ldots, N/M)^T$.

Figure 4-8 shows the actual distribution of cumulative probability measurements for all formants in the training data. The top panel is a Q-Q plot, in which the 627 sorted data points are plotted against 627 equally spaced quantiles from a uniform distribution. The bottom panel is a histogram with $M = 50$ bins. A T^2 test using the covariance matrix in equation 4.10 shows that this histogram is significantly different from a uniform distribution ($F = 2.209, p \leq 0.01$). The most important difference seems to be that the distribution is tilted: actual formants are likely to be lower in frequency than the formants predicted by the *a posteriori* formant models.

Figure 4-9 shows the distribution of cumulative probability measurements for all formants in the test data. As shown by the large bar in the leftmost histogram bin, about 10% of the test formants fall within the lowest 2% of the *a posteriori* formant probability distributions; in other words, in about 10% of the test tokens, the human transcriptions are lower in frequency than any value considered possible by the HMM. Informal analysis of the data indicates that these overshoot errors are found more or less equally in all three formants. The remaining 90% of the cumulative probability measurements are relatively uniform, although still significantly different from a true uniform distribution.

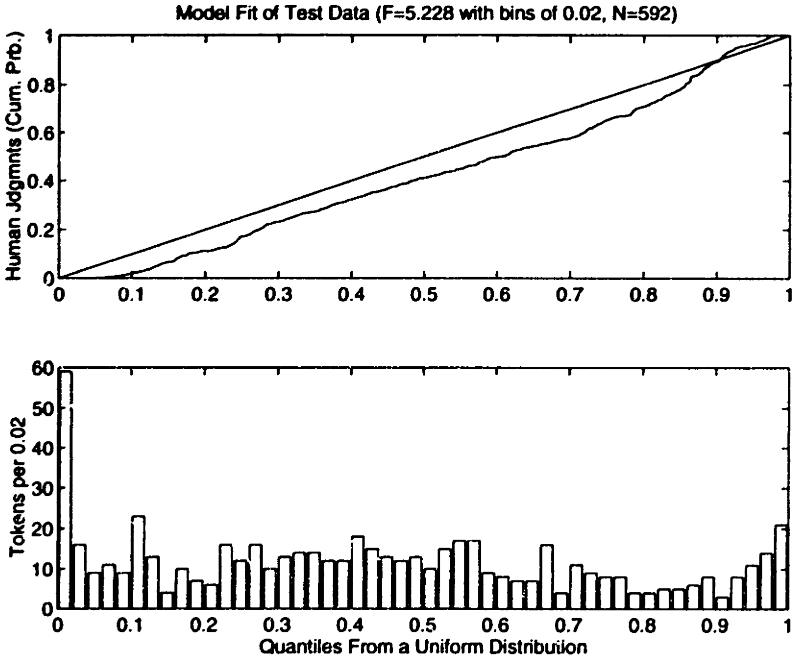


Figure 4-9: Cumulative probability representations of formants measured by two human judges on an independent test set (F1, F2, and F3). Deviation from a uniform distribution indicates that the formant tracker is not adequately predicting variation in the new corpus.

4.5 Conclusions

The HMM formant tracker developed in this chapter, despite its simplicity, provides useful token-by-token models of the formant measurement uncertainty.

The uncertainty models were found to be significantly different from the distribution of true formant measurements. Specifically, the uncertainty models predicted formants which were higher in frequency, on average, than the formants measured by human transcribers. Although this effect is significant, it is not large: only 10% of the test measurements were below the range predicted by the formant tracker.

It should be noted that the fact that we were able to perform this analysis at all is proof of the value of Bayesian formant tracking. The performance of formant tracking algorithms is often rated in terms of an aggregate error rate, but apparently no other formant tracking algorithm has ever attempted to report its own measurement uncertainty for each token, individually.

Chapter 5

Classification of Place

In section 1.1, we argued that perfect formant and burst spectral measurements can be used to identify the place of a stop consonant. This chapter describes one modeling experiment and two classification experiments which explore the effect of measurement error on classification.

The rule-based measurements developed in chapter 3 are explored here in some detail. First, a model is developed which predicts the effect of measurement error on the error rate of a linear discriminant analysis (LDA) classifier, and this model is tested using a database with both manual and automatic measurements. Second, the measurements are evaluated in a context-dependent LDA classifier, in which the test data consists of the entire TEST subdirectory of TIMIT (Zue et al., 1990). Context-dependent LDA classification of place is shown to be 84% correct (in vowel and glide contexts) using automatic measurements of formant frequency and burst spectral shape.

In the third experiment in this chapter, the stochastic formant model (with some modification of the design in chapter 4) is tested in context-dependent classification of voiced stops from the TEST subdirectory of TIMIT. The stochastic formant model classifies consonants on the basis of formant frequency and amplitude in the first 50ms following release, and results in an 83% correct classification rate in vowel and glide contexts.

Most of the experiments in this chapter report the result of classifying the place of articulation of voiced and voiceless stop consonants. Section 5.2.2 reports linear discriminant classification of nasal releases, using automatic formant frequency measurements. In the HMM classification experiment in section 5.3, only voiced stops are classified.

5.1 Effect of Measurement Error on the Linear Discriminant

This section discusses the effect of measurement error on linear discriminant analysis. The database to be analyzed in this section consists of manual and automatic formant and burst measurements of 131 voiced and voiceless stops in vowel and glide context, extracted from the Error Modeling database (appendix A.2). The difference between LDA classification of human and automatic measurements is taken to be the effect of measurement error. Attempts are made to predict the effect of measurement error on the linear discriminant using an aggregate error model.

LDA is used in these experiments because the algorithm is fast, and because it provides a simple parametric form which can easily be combined with a quantitative error model. It should be emphasized, as noted in chapter 1, that LDA is known to be suboptimal

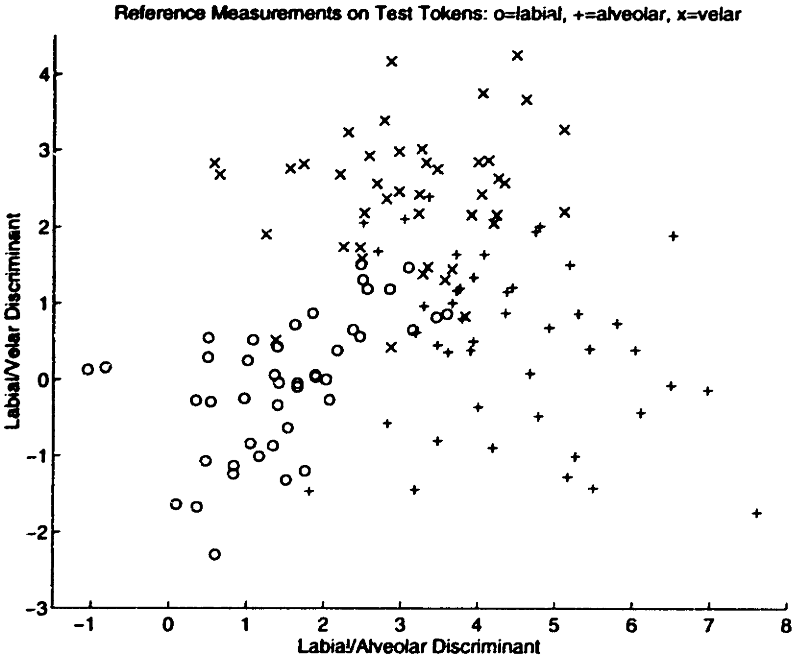


Figure 5-1: LDA summary of ineasurements by human judges on 131 stop releases. The abscissa is a linear discriminant trained (on independent training data) to represent the labial/alveolar distinction; the ordinate is a discriminate trained to represent the labial/velar distinction.

in phonetic classification tasks. The experiment in this section demonstrates 89% correct LDA classification of exactly the same numbers (manual formant and burst measurements extracted from the Error Modeling database) which are separated with approximately 95% accuracy in section 1.1. There are two differences between these two experiments: the type of the classifier (LDA in this section, vs. a knowledge-based design in section 1.1), and the training of the classifier (the classifier in section 1.1 is trained on the test data, while the classifier in this section is trained on an independent data set). Without another experiment, it is impossible to separate these two effects. Based on a comparison of these two experiments, then, we are limited to the rather weak claim that LDA classification increases error rate in this experiment by *up to* six percentage points.

5.1.1 Measuring the effect of error

In order to test models of the linear discriminant error, three context-independent binary LDA classifiers were trained on manual transcriptions of the KB Train database (appendix A.1). Each classifier represents one binary place distinction: labial/alveolar, labial/velar, or alveolar/velar. The labial/alveolar and labial/velar discriminants make use of nine measurements: two burst spectral amplitudes and a peak count, and measurements of the first three formants at 20ms and 50ms after the consonant release. The alveolar/velar distinction makes use of a tenth measurement, the burst front cavity resonance, which is not available to the other two discriminant vectors.

Classifiers were tested using 131 stop releases in vowel right context from the Error Modeling database (appendix A.2). Figure 5-1 shows the labial/alveolar and labial/velar discriminant representations of manual transcriptions of this database. Figure 5-2 shows

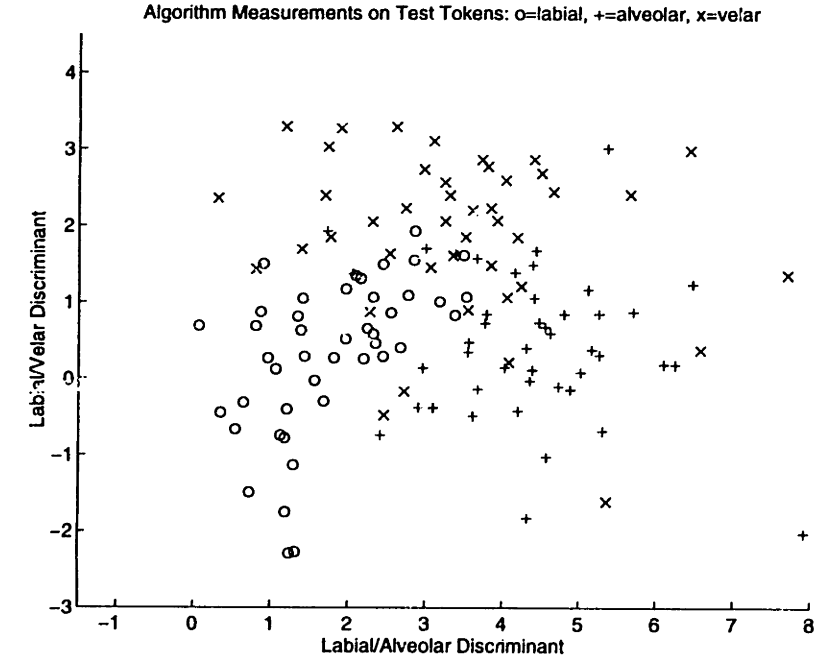


Figure 5-2: LDA summary of automatic measurements of burst amplitude, burst peak count, and formant motion at 131 stop releases. The abscissa and ordinate are computed using the same linear discrimant coefficients applied to the data in figure 5-1.

Spoken	Manual Measurements				Automatic Measurements			
Place	lips	blade	body	unknown	lips	blade	body	unknown
lips	80	0	11	9	66	2	20	11
blade	2	95	2	0	0	82	11	7
body	2	0	93	5	2	7	81	9

Table 5.1: Percentage confusion matrices showing classification of vowel context stops in the Error Modeling database. The first confusion matrix shows classification using manual transcriptions; the second shows classification using automatic measurements.

the same two-dimensional representation of an automatic transcription of the same data, using the algorithms developed in chapter 3.

Table 5.1 describes the classification performance of the LDA classifier, using the measurements shown in figures 5-1 and 5-2. This table uses a round-robin classification strategy: each token is separately classified by all three binary classifiers, and the three classifiers vote to determine the final place label. If there is no majority, the token is marked as "unknown."

Classification using manual measurements in table 5.1 is relatively good, at 89% correct classification. This is comparable to the results reported by Lamel (1988), who found that manual transcriptions of formants, burst measures, and voice onset time could be used to classify stops with roughly 90% accuracy. Classification using automatic measurements is worse by 12-14% for every place of articulation, at an average of 76% correct classification.

5.1.2 The form of the discriminant error term

Suppose that $\mathbf{x} = \mathbf{x_0}+\mathbf{e}$ is a possibly erroneous, known measurement of an unknown formant frequency $\mathbf{x}$. A linear discriminant statistic ξ is a linear combination of the elements of $\mathbf{x}$ which is intended to be useful for classification:

$$\xi = \mathbf{z}'\mathbf{x} = \mathbf{z}'\mathbf{x_0} + \mathbf{z}'\mathbf{e} \tag{5.1}$$

This discussion will assume that the coefficients $\mathbf{z}$, and the linear discriminant classification threshold, have been previously trained on a data set which does not contain $\mathbf{x_0}$, and that the coefficients and threshold can be considered fixed.

If the mean $E[\mathbf{e}]$ and covariance matrix $\mathbf{C}_e$ of $\mathbf{e}$ are known, the mean and variance of the linear discriminant error term, $\epsilon = \mathbf{z}'\mathbf{e}$, are easily calculated.

$$E[\epsilon] = \mathbf{z}'E[\mathbf{e}], \quad \sigma_\epsilon^2 = \mathbf{z}'\mathbf{C}_e\mathbf{z} \tag{5.2}$$

If $\mathbf{e}$ is Gaussian, ϵ is also Gaussian, and as such, is completely described by its mean and variance.

If $\mathbf{e}$ is mixture Gaussian, ϵ is also mixture Gaussian. Specifically, if each element e_i of the measurement error is well modeled by n_i mixture elements, then the number of mixtures required to fully model ϵ is

$$n_\epsilon = \prod_i n_i \tag{5.3}$$

The n_ϵ different mixture elements describing ϵ can be conveniently indexed by the vector $\mathbf{k} = (k_1, k_2, \ldots)$. Suppose that mixture k_i of element e_i from the error vector is described by a mean $\mu_i(k_i)$ and a variance $\sigma_i^2(k_i)$, that the probability of choosing this mixture is $P_i(k_i)$, and, finally, that all of the elements of $\mathbf{e}$ are independent. In this case, the mixture elements describing ϵ have the following mean, variance, and probability of occurrence:

$$\mu_\epsilon(\mathbf{k}) = \sum_i z_i\mu_i(k_i), \quad \sigma_\epsilon^2(\mathbf{k}) = \sum_i z_i^2\sigma_i^2(k_i) \tag{5.4}$$

$$P_\epsilon(\mathbf{k}) = \prod_i P_i(k_i)$$

The above expression for $P_\epsilon(\mathbf{k})$ can be used in more general, nonparametric models of the discriminant error. For example, if each element e_i of the error is known to be large with probability $P_i(+)$ independent of every other element e_j, and if ϵ is known to be large whenever any of the individual measurement errors is large, then ϵ is large with probability

$$P_\epsilon(+) = 1 - \prod_i(1 - P_i(+)) \approx \sum_i P_i(+) \tag{5.5}$$

5.1.3 Estimating confusion matrices from the discriminant error

Given quantitative models of the manual measurements, and of the measurement error, it is possible to predict a confusion matrix for the automatic measurements. This section demonstrates prediction of the automatic measurements confusion matrix using Gaussian models of each phoneme class, and of the measurement error.

Binary discriminant classification of two Gaussian distributions is shown schematically

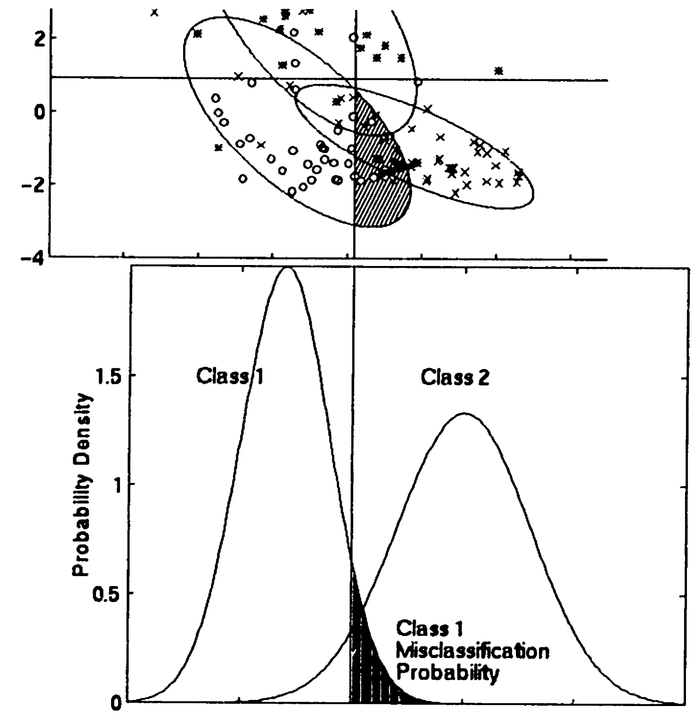


Figure 5-3: Binary classifier. If both distributions are Gaussian as shown, the shaded area of the category 1 distribution is the probability that a category 1 token will be misclassified.

in figure 5-3. In this figure, multivariate Gaussian distributions have been reduced to univariate Gaussians by multiplication with the discriminant vector $\mathbf{z}_{12}$. The univariate measurements $\xi = \mathbf{z}_{12}'\mathbf{x}$ are classified in category 1 if $\xi < \theta_{12}$, and category 2 if $\xi > \theta_{12}$, where the threshold θ_{12} and discriminant vector $\mathbf{z}_{12}$ have been previously trained on an independent data set. If the tokens in category 1, $\xi \subset G_1$, are distributed normally with mean $\mu_{\xi 1}$ and variance $\sigma_{\xi 1}^2$, the probability that a token from G_1 will be classified as G_2 is

$$P_{12}(\widehat{G_2}|G_1) = \int_{\frac{\theta_{12}-\mu_{\xi 1}}{\sigma_{\xi 1}}}^{\infty} N(t)dt \qquad (5.6)$$

where $N(t)$ is the unit normal distribution.

In a round-robin classification scheme, the classifier in figure 5-3 is given tokens from all three groups, G_1, G_2, and G_3, and asked to label each token as G_1 or G_2. Tokens from G_3 are necessarily misclassified, since the classifier in figure 5-3 doesn't know about G_3. We can characterize the behavior of the classifier in figure 5-3 with respect to G_3 tokens using the complementary probabilities $P_{12}(\widehat{G_1}|G_3)$ and $P_{12}(\widehat{G_2}|G_3)$:

$$P_{12}(\widehat{G_1}|G_3) = \int_{-\infty}^{\frac{\theta_{12}-\mu_{\xi 3}}{\sigma_{\xi 3}}} N(t)dt, \quad P_{12}(\widehat{G_2}|G_3) = \int_{\frac{\theta_{12}-\mu_{\xi 3}}{\sigma_{\xi 3}}}^{\infty} N(t)dt \qquad (5.7)$$

If the three binary classifiers are assumed to be independent, the entries in a confusion matrix can be calculated by multiplying the classification probabilities of each binary

Spoken	Homoskedastic Model			
Place	lips	blade	body	unknown
lips	54	7	18	21
blade	7	76	5	12
body	6	6	73	15
	Heteroskedastic Model			
lips	60	5	15	20
blade	4	83	4	9
body	4	5	78	13

Table 5.2: Automatic measurement confusions predicted based on statistics of the manual measurements, and of the measurement error, using homoskedastic and heteroskedastic models.

classifier, thus

$$\Pr(\widehat{G_1}|G_1) = (1 - P_{12}(\widehat{G_2}|G_1))(1 - P_{13}(\widehat{G_3}|G_1)) \tag{5.8}$$

$$\Pr(\widehat{G_2}|G_1) = P_{12}(\widehat{G_2}|G_1)P_{23}(\widehat{G_2}|G_1) \tag{5.9}$$

$$\Pr(\widehat{G_3}|G_1) = P_{13}(\widehat{G_3}|G_1)P_{23}(\widehat{G_3}|G_1) \tag{5.10}$$

and the probability of a token being marked unclassifiable is the complement of the three probabilities shown.

Using equations 5.8 to 5.10, it is possible to predict the classification of one set of measurements using information from another set. For example, if the means and variances of the automatic discriminant measurements ξ_a are related to the means and variances of the manual measurements ξ_m by the addition of measurement error,

$$E[\xi_a] = E[\xi_m] + \mu_\epsilon, \quad \sigma^2_{\xi_a} = \sigma^2_{\xi_m} + \sigma^2_\epsilon \tag{5.11}$$

then classification of the automatic measurements can be predicted from classification of the manual measurements.

Table 5.2 shows confusion matrices for the automatic measurement set, predicted using homoskedastic and heteroskedastic models of the measurement error. In computing this table, the reference measurement vector $\mathbf{x}_0$ was assumed to have full covariance, while the measurement error $\mathbf{e}$ was assumed to have diagonal covariance. The mean and covariance of $\mathbf{x}_0$ were measured directly on the 131 stop releases analyzed in this section. The mean and variance of the measurement error were copied from models developed in section 3.4 based on the full Error Modeling database, a superset of the current test data. The correlation between measurements and measurement error in the heteroskedastic model was copied from table 3.5. Extending equation 5.11 for a heteroskedastic error model and a full-covariance measurement vector required a somewhat lengthy but straightforward derivation, which is not reproduced here.

The classification of automatic measurements is much more accurately predicted by a heteroskedastic than a homoskedastic error model. Apparently, correlations between the measurement values and measurement error, reported in section 3.4.5, are carried through into the linear discriminant statistics. In fact, the heteroskedasticity of the discriminant statistics is visible in figures 5-1 and 5-2. In these figures, the automatic measurement distributions are more compact than the manual measurement distributions, implying that

Error Model	Classification of tokens as			
	lips	blade	body	unknown
Homoskedastic	7.7	4	5.3	7
Heteroskedastic	4	2	5	5

Table 5.3: Average magnitude difference, in each column, between the classification of automatic measurements, shown in table 5.1, and the predicted classification using homoskedastic and heteroskedastic error models, shown in table 5.2.

the measurement algorithm is selectively increasing low amplitude discriminant statistics, and decreasing high amplitude statistics, as is characteristic of heteroskedastic error.

Section 3.4.5 concludes that the most heteroskedastic measurements are the two burst amplitudes, and the frequency of F1 near onset. According to the speech production theory sketched in chapter 2, these three measurements are all useful in distinguishing labial from lingual tokens. This prediction from theory is empirically confirmed by the linear discriminant vectors trained in section 5.1.1: the labial/alveolar and labial/velar discriminant vectors weight burst amplitude four times as heavily, and F1 onset ten times as heavily, as does the alveolar/velar discriminant vector.

Since the labial/alveolar and labial/velar discriminants weight burst amplitude and F1 onset heavily, the classification of tokens as labial (that is, the first column of the confusion matrix) should be affected more by the difference between homoskedastic and heteroskedastic models than the other columns. Table 5.3 shows that this is in fact the case. In this table, the actual classification of automatic measurements, in table 5.1, has been subtracted from the predictions made by homoskedastic and heteroskedastic error models, shown in table 5.2, and the magnitude differences have been averaged for each column of the confusion matrix. As shown, the average error of the homoskedastic model in predicting the actual classification performance in table 5.1 is worst in the first column, which reports the percentage of stops classified or misclassified as labial. The first column also shows the greatest total improvement between the homoskedastic and heteroskedastic models.

5.2 Classification using Rule-Based Measurements

This section describes an experiment in which the rule-based measurements described in chapter 3 are optimized in order to give the best possible context-dependent linear discriminant classification of consonant place. For each right context, speaker gender, and consonant class, specific parameters of the measurements are optimized using all of the relevant consonant releases in the TRAIN subdirectory of TIMIT, excluding those which cross a word boundary. The resulting measurements are then tested, again using linear discriminant classification, on all of the relevant consonant releases in the TEST subdirectory of TIMIT, again excluding those which cross a word boundary.

The TIMIT database is transcribed with a sampling period of five milliseconds, which is close to the duration of a short frication burst. In order to ensure measurement of the entire frication burst, the release times of all stops in the training and test databases were re-transcribed with a sampling period of one millisecond. Nasal releases were not re-transcribed.

	tongue height					
tongue position	syllable on-glide	advanced tongue root	lax high	lax low	constricted pharynx	reduced
front	/y/	/iy,ux/	/ih/	/eh/	/ae/	/ix/
back	/w/	/uw/	/uh/	/ah/	/aa/	/ax/
retroflex	/r/	/er/				/axr/
lateral	/l/	/el/				

Table 5.4: Context-dependent measurement parameters were optimized independently for each combination of a speaker gender, a consonant manner class, and a right context. The 18 analyzed right contexts are shown here. Phoneme notation is ARPABET, as used in the TIMIT database.

5.2.1 Optimizing free parameters

For each right context, speaker gender, and consonant class, specific parameters describing the measurement algorithms were trained, in order to minimize classification error, on all applicable syllables from the TRAIN subdirectory of TIMIT.

Grouping similar contexts for training

Parameters describing each measurement algorithm were optimized in order to minimize error in the classification of place using a context-dependent LDA classifier. Each set of context-dependent measurement parameters was optimized independently for 108 context cells, where the 108 cells included every possible combination of 2 genders, 3 consonant classes, and 18 right contexts. The 18 right contexts are given in table 5.4.

For many of the context cells defined in this way, the TRAIN subdirectory of TIMIT does not contain enough tokens to allow reliable training of the classifier. In order to improve parameter training, sparse cells were supplemented by tokens having similar right contexts, in order to maintain a minimum of 5 tokens of each consonant per linear discriminant coefficient. Similarity of right contexts was measured separately for each consonant and each gender. The similarity metric was a simple Euclidean distance metric, measured between average formant frequencies (linear scale) and amplitudes (logarithmic scale) at 10ms intervals during the first 80ms following consonant release.

Frequency band definitions for burst spectral measurements

Chapter 3 defines four burst spectral measures in terms of three frequency bands: a low-frequency band, a high-frequency band, and a front cavity resonance spectrum. In the experiment described here, the low and high edge frequencies for each of these bands were optimized for each context cell, independently, in order to minimize classification error using a context-dependent LDA classifier.

Band edge frequencies were optimized using a discrete gradient search algorithm. Each edge frequency was initialized to the frequency specified in section 3.3.3, and the initial frequency values were used to estimate an LDA classification error using the training data. The frequency parameters were then modified iteratively, in steps of one DFT bin (62.5Hz), with the single frequency parameter providing the greatest classification improvement modified at each step. During optimization, frequency parameters were not allowed to leave the bands specified in table 5.5.

Frequency Bands in the Burst Spectrum			
low-frequency amplitude	$(l_1 \in [880, 1440])$	to	$(h_1 \in [2250, 3440])$
high-frequency amplitude	$(l_2 \in [2250, 3440])$	to	$(h_2 \in [5690, 7440])$
front cavity resonance	$(l_3 \in [880, 1440])$	to	$(h_3 \in [5690, 7440])$

Table 5.5: The six band-edge frequencies shown were optimized, within the ranges shown, in order to minimize consonant place classification error in context-dependent LDA classifiers.

Timing of formant frequency samples

The formant smoothing algorithm described in chapter 3 was used, without modification, to smooth measurements of F1, F2, and F3 between 0 and 80ms after release (step 5ms).

Samples were then chosen from each formant frequency track, for each context cell, in order to minimize the context-dependent LDA classification error. Voiced and unvoiced stops were first classified using the optimized burst spectral measurements, and then formant samples were added, beginning with the single sample which most decreased classification error. Formant selection for nasal consonant classes was similar, but with no burst spectral measurements to initialize the algorithm.

In this way, the classifier for each context cell was assigned five formant samples, including at least one from each of the three formant tracks.

5.2.2 Linear discriminant classification

Burst and formant measurements, trained as described above, were combined using linear discriminant analysis in order to classify the place of consonant releases in the TEST subdirectory of TIMIT.

The test database consisted of all stop and nasal releases in the 18 right contexts listed in table 5.4, in the TEST subdirectory of TIMIT, which do not span a word boundary (that is, the consonant is not word-final): a total of 2804 stops, and 1461 nasals. The following two subsections discuss the classification of stop releases and the classification of nasals, respectively.

Classification of stops

Table 5.6 describes the results of context-dependent LDA classification of stops. Classification scores for male and female utterances of both voiced and unvoiced stops have been pooled. The three retroflex contexts have been pooled under the heading "retroflex," the two lateral contexts under the heading "lateral," and all other contexts under the heading "vowels and glides." In the table, labial, alveolar, and velar stops have been labeled with the name of the primary articulator: lips, tongue blade, and tongue body, respectively.

In classification, the three places of articulation were assumed to be *a priori* equally likely, despite the obvious differences in number of tokens (NT). In vowel and glide contexts, classification using formant measures is 72% correct, classification using burst measures is 80% correct, and classification using both formant and burst measures is 84% correct.

Regardless of the measurements used, alveolar and velar stops are more likely to be confused with each other than with labial stops. This is somewhat surprising, since labial classification is significantly worse than velar or alveolar classification in the context-independent classifier of section 5.1 (table 5.1). Apparently, context information improves classification of labial stops, but has little effect on classification of alveolar and velar stops.

Linear Discriminant Classification of Stops

Spoken			Formants			Burst			Both		
V	C	NT	lips	blade	body	lips	blade	body	lips	blade	body
Vowels	lips	771	80%	13		83			89		
and	blade	632	14	70	16	12	77	11		82	
Glides	body	602	15	20	65		12	79		13	81
Retroflex	lips	209	76		16	86			86		
	blade	194		71	22	23	61	16	14	70	16
	body	128	31	23	45		14	81		16	81
Lateral	lips	127	75		20	80	16		90		
	blade	30	33	40	27	27	67		23	70	
	body	111	39	13	49			86			88

Table 5.6: Confusion matrices using context-dependent formant and burst spectral measurements, categorized using a context-dependent LDA classifier. Stop tokens only, from the TEST subdirectory of TIMIT. All entries of 10% or less have been omitted.

Linear Discriminant Classification of Nasals

Spoken			Formants		
V	C	NT	lips	blade	body
Vowels	lips	576	78%	16	
and	blade	563	16	71	
Glides	body	50		36	58
Retroflex	lips	53	72		17
	blade	70		47	39
	body	9	22	44	33
Lateral	lips	25	28	32	40
	blade	95	20	51	29
	body	20	30	45	25

Table 5.7: Confusion matrices: linear discriminant classification of nasal consonants, using automatic formant measurements. TEST subdirectory of TIMIT. All entries of 15% or less have been omitted.

Formant classification degrades significantly in both lateral and retroflex contexts: the neighboring lateral or retroflex sound apparently dominates the evolution of formant frequencies, reducing the distinction between different places of stop articulation. Classification scores which make use of burst spectral measures, on the other hand, are not significantly degraded: classification in retroflex context using both burst and formant measures is 79% correct, and classification in lateral context is 87% correct.

Classification of nasals

Table 5.7 describes the results of context-dependent LDA classification of nasal releases.

Velar nasal releases (/ng/) are quite poorly recognized. There are not very many of these. The phoneme /ng/ can never begin a syllable in English; since releases which span a word boundary were excluded from both training and test data, all of the /ng/ releases in both data sets come from word-internal syllable boundaries. Automatic combination of

Humans, Two Syllables (Lamel, 1988)

Spoken	Vowel context				Semivowel context			
C	NT	lips	blade	body	NT	lips	blade	body
lips	219	99.4%	0.5		92	98.9	C.7	0.4
blade	210		99.2	0.6	59		99.6	
body	204	0.3		99.6	97			99.9

Table 5.8: Classification of consonant place by human listeners, given full closure and release information (Lamel, 1988). Entries below 0.3% have been omitted.

Classification, Human Listeners (Nossair and Zahorian, 1991)

Spoken		Burst+50ms			Burst+Transition			Burst+Full Vowel		
C	NT	lips	blade	body	lips	blade	body	lips	blade	body
lips	756	97.9%		2.0	99.2			99.2		
blade	756	1.4	93.7	5.0	1.1	95.2	3.7		96.4	3.2
body	756	1.9	1.4	96.8			99.0			99.0

Table 5.9: Classification of consonant place by human listeners, given release waveforms of varying length (Nossair and Zahorian, 1991). Entries less than 1% have been omitted.

similar context cells was used to guarantee that no classifier was ever trained with fewer than 25 /ng/ release tokens, but in many cases the number of labial and alveolar tokens was much larger than 25. Measurement definitions were trained to minimize a total error rate; since there are so few velar tokens, it is possible that the measurement parameters were automatically adjusted to minimize labial and alveolar errors, with little concern for velar errors.

Classification of all nasal releases in vowel and glide context is 74% correct, which is quite close to the classification rate of stops if only formant frequencies are used. In retroflex and lateral contexts, nasal classification degrades more than stop classification, to about 56% correct in retroflex contexts, and 43% correct in lateral contexts.

5.2.3 Comparison to the performance of human listeners

It is difficult to compare the results of the current study to the classification ability of humans, since human listeners can identify consonant place in nonsense words with better than 99% accuracy. Table 5.8, for example, describes the ability of human listeners to classify stop place of articulation, as measured by Lamel (1988) (voiced and unvoiced cognates listed in her confusion matrices have been combined). In this study, listeners were given an entire natural vowel-consonant-vowel sequence, extracted from a continuous read sentence, and asked to identify the consonant.

The ability of listeners to make use of burst information exclusively, or of burst and formant information together, can be measured by gating the signal, so that listeners hear only the desired portion of a stop release. Table 5.9 compares the ability of listeners to classify stops given three successively larger sections of a stop release, extracted from a monosyllabic word recorded in isolation, as measured by Nossair and Zahorian (1991) (voiced and unvoiced cognates listed in their confusion matrices have been combined). Given a full release syllable through the end of the following vowel, listeners were able to classify stop place with 98.2% accuracy. Given only the burst and the formant transition, classification accuracy dropped very slightly to 97.8%. Finally, given only a 50ms waveform beginning at release, listeners were only able to correctly identify 96.1% of the stops.

In the 50ms condition, Nossair and Zahorian report that unvoiced stops were classified much more accurately (93.6%) than voiced stops (86.4%), indicating that voiced stops may be more difficult to classify on the basis of the burst and formant onsets than are unvoiced stops. Since the burst was apparently less effective for classification of voiced stops, they tried an additional experiment in which subjects classified voiced stop syllables, including the entire vowel, but with the frication burst removed. Subjects were completely unable to classify stops without the burst: average classification was only 74.5% correct. Apparently, a voiced stop can only be classified given both burst and formant transition information.

5.2.4 Effect of right context

The effect of retroflex and lateral right contexts has been discussed above, in connection with tables 5.6 and 5.7. In those tables, we saw that classification of stops or nasals on the basis of formant frequencies degrades significantly in retroflex or lateral context, but that stops can still be classified relatively well given burst spectral information. Given burst spectral information, there is a significant difference between retroflex and lateral contexts: in retroflex contexts, confusions between alveolar and velar stops increase, while in lateral context, confusions between alveolar and labial stops increase.

The classifier performance can be compared to the human performance reported by Lamel, and given in table 5.8. In this study, all phonotactically possible combinations of a stop plus a semivowel /l,r,w/ were presented to listeners (i.e. no alveolars in lateral context), along with all of the affricates, and listeners were asked to identify the phoneme. Affricates and stops identified as affricates have been omitted from table 5.8. The error pattern on the right-hand side of table 5.8 shows increased misclassification of labial stops, which is similar to the error pattern of the LDA classifier in lateral context. The increased alveolar/velar confusion of the LDA classifier in retroflex context, however, was not observed in Lamel's data; instead, listeners in Lamel's study tended to misclassify alveolars as affricates in retroflex context.

Another study which has published an analysis of human stop identification errors as a function of right context was produced by Winitz et al. in 1971. In this study, subjects were presented with unvoiced stop releases in /aa,iy,uw/ context, and asked to identify the stop. In the first listening condition, subjects were provided with only the unvoiced portion of the release; in the second condition, subjects were given the unvoiced portion, plus 100ms of the following vowel. The results are summarized in table 5.10, and compared to the classification performance of the LDA classifier in similar contexts (tense, lax, and glide contexts are grouped together in the LDA classifier results).

The classification scores reported by Winitz et al. are significantly lower than the scores reported by Nossair and Zahorian, despite the fact that the two windows used by Winitz et al. were apparently longer than the 50ms and burst+transition windows used by Nossair and Zahorian. The difference between these two studies can probably be attributed to distortion of the tokens used by Winitz et al., which were recorded, gated, re-recorded, and played back to subjects using analog equipment.

Despite the unreasonably high average error rates in the Winitz et al. study, some of the distributions of errors may be roughly correct. For example, Ohala has argued (1989) that the tendency of listeners to identify labial and velar stops as alveolar in high front (/iy/) context is correlated with cross-language patterns of historical sound change, in which labials

Linear Discriminant Classification, All Stops

Spoken			Burst			Burst and Formants		
V	C	NT	lips	blade	body	lips	blade	body
LB	lips	126	90			97		
/aa,	blade	68	16	75			82	
ah/	body	176		14	77		15	80
HF	lips	311	82			89		
/iy,ih,	blade	224		79	13		84	
y,ux/	body	120		16	74		19	74
HB	lips	47	72		19	81		17
/uw,	blade	84	11	76	13		83	12
uh,w/	body	84		13	83			87

Human Classification (Winitz et al., 1971)

/aa/	/p/	52	55	32		88		
	/t/	52	22	73			75	
	/k/	52	24	46	30		37	49
/iy/	/p/	52	64	19	16	55	34	
	/t/	52	20	65			85	
	/k/	52	32	46	22	27	32	41
/uw/	/p/	52	69	21		72		16
	/t/	52	20	64			80	
	/k/	52	34	23	43	27	23	50

Table 5.10: Percentage classification of place as a function of right context. LB=low back vowels, HF=high front vowels and glides, HB=high back vowels and glides. Entries below 10% (LDA classifier) and 15% (human subjects) have been omitted.

and velars in this context tend to become tongue blade consonants. This can be compared to the pattern of errors of the LDA ciassifier. The classifier frequently mis-labels velars as alveolar in high front context, but seems to have little trouble with labials.

Both the LDA classifier and the human listeners achieve their best classification scores in low back contexts (/aa/ and /ah/). In this context, the most likely confusion, for both the LDA classifier and human listeners, is the labeling of a velar stop as alveolar.

In high back context, the linear discriminant classifier tends to misclassify labial stops as velar, while human listeners tend to misclassify velar stops as labial. It is possible that this difference is caused by differences in the number of tokens. The measurement training data contains 243 velar releases in these contexts, compared to only 99 labial releases, so it is possible that the minimum-error training algorithm adjusted the measurement parameters to minimize velar errors at the expense of additional labial errors.

5.3 Stochastic Formant Modeling

This section describes an experiment in which two-level phoneme-dependent HMM formant models, described in chapter 4, are trained based on automatic transcriptions of all of the voiced stop tokens in the TRAIN subdirectory of TIMIT, and tested using spectra from all relevant tokens in the TEST subdirectory, excluding releases which span a word boundary. Al' release times are given by the transcription with a precision of one millisecond, as described at the beginning of section 5.2.

5.3.1 Modifications to the model

In order to allow use as a phoneme recognizer, the HMM formant trackers described in chapter 4 are modified in several ways.

Implementation of a two-level search space

First, the two-level search space shown in figure 4-1 is implemented. Phoneme classification proceeds according to a maximum likelihood rule: the possible phoneme models are tested, and the model which is most likely to have produced the observed spectrogram is chosen as the correct label.

The probability of an observed spectrogram, given a particular phoneme model λ, can be written

$$\Pr(\mathbf{O}(0),\ldots,\mathbf{O}(T)|\lambda) = \sum_{F1(t)} \sum_{F2(t)} \sum_{F3(t)} \Pr(\mathbf{O}(0),\ldots,\mathbf{O}(T),F_1(t),F_2(t),F_3(t)|\lambda) \tag{5.12}$$

where the sum on the right-hand side can be carried out at any time t.

If the formant frequencies are assumed independent, the right-hand side of equation 5.12 can be factored into three independent formant probabilities, which can be computed using the forward-backward algorithm, as discussed in section 4.3.1:

$$p(F_n(t),\mathbf{O}(0),\ldots,\mathbf{O}(T)|\lambda) \equiv \alpha_n(F_n(t),t)\beta_n(F_n(t),t) \tag{5.13}$$

Recoding the output model in terms of convex peaks

In the classification experiments described here, the stochastic formant model was trained and tested several times on the entire TRAIN subdirectory of TIMIT, using various combinations of parameters. In order to expedite retraining of the model, each training and test spectrogram was reduced to a representation consisting of a list of convex peak frequencies and their amplitudes, and this representation was pre-computed for all tokens, and stored on disk.

This simplified spectral representation entails several modifications to the DFT output model described in section 4.1.2.

First, the spectral convexity $C(f,t)$ is quantized to a binary "peak/not-peak" distinction. The half-Gaussian model of spectral convexity described in section 4.1.2 is eliminated. Instead, the binary convexity of a formant is modeled as a simple Bernoulli random process, in which the probability of a formant $F_n(t)$ landing on a convex spectral peak at each time t is a trainable parameter, $P_n(t)$,

$$P_n(t) \equiv \Pr(C(F_n,t) = 1) \tag{5.14}$$

The logistic spectral amplitude model described in section 4.1.2 is maintained without change when evaluating the frequency of a convex spectral peak. At frequencies between spectral peaks, the output representation does not specify a spectral amplitude $A(f,t)$. In the experiments described in this chapter, the forward-backward algorithm automatically assigns the amplitude probability a low constant value k for all frequencies between spectral peaks,

$$P(A(f,t)|C(f,t) = 0) = k \tag{5.15}$$

The value of this fixed output probability is chosen to be low enough to cause the maximum *a posteriori* probability of a formant to always occur at the frequency of a convex peak, unless the nearest convex peak is several standard deviations away from the expected formant frequency.

Temporal correlation in the output model

The normalized amplitude of a formant may be temporally correlated, just as the frequency is. In the experiments described in this chapter, the temporal correlations of formant amplitudes (normalized by the spectral energy) are modeled by a trainable correlation coefficient $\rho_{An}(t)$, analogous to the formant frequency correlation parameter $\rho_{Fn}(t)$ described in section 4.1.2. In training, $\rho_{An}(t)$ was generally found to be much less than $\rho_{Fn}(t)$, for each formant; typical values of $\rho_{An}(t)$ vary between 0.0 and 0.5. A few informal experiments indicated that classification with $\rho_{An}(t)$ is not significantly better or worse than classification without it.

Because of the large quantity of training data, all model parameters (including means, variances, correlation coefficients, and Bernoulli probabilities) were trained independently for every time t.

Additional spectral measurements

In order to improve place classification, the model in this chapter uses two additional spectral measurements which are not described in chapter 4.

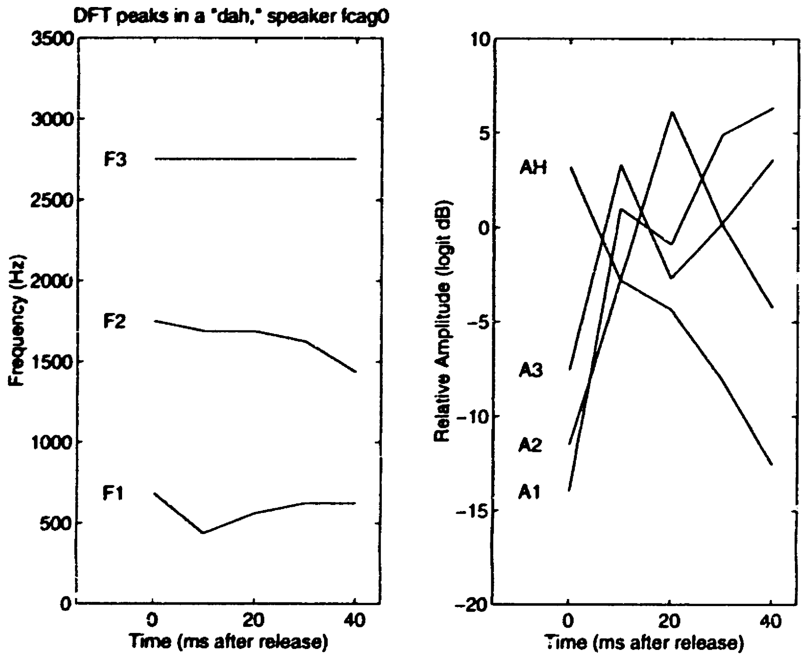


Figure 5-4: Example of three convex peak traces at the release of a /d/ by a female speaker. The left-hand plot shows the convex peak frequencies closest to expected values of F1, F2, and F3. The right-hand plot shows the amplitudes of these three peaks, and the amplitude AH of the largest peak higher in frequency than F3.

First, the overall spectral energy $E(t)$ is modeled as an independent output variable. The spectral energy in decibels is modeled as a Gauss-Markov process, independent of any of the formants. The mean, variance, and temporal correlation of the energy as a function of time are trained separately for each context-dependent phoneme model.

Second, the amplitude of the largest spectral peak above F3 is measured for each possible value of F3, and the probability distribution of this high-frequency amplitude measure, $A_{f\geq}(f)$, is used in the output model for F3. For example, at frequencies $F_3(t)$ which are convex peaks,

$$\Pr(\mathrm{O}(t)|F_3(t),\lambda) = P_3(t)\Pr(A(F_3,t)|F_3(t),\lambda)\Pr(A_{f\geq}(F_3,t)|F_3(t),\lambda) \tag{5.16}$$

where $P_3(t)$ is the trained probability of F3 landing on a convex peak, as discussed in section 5.3.1. The $A_{f\geq}$ measurement was added to allow some modeling of the high-frequency tilt of an alveolar burst spectrum. In informal experiments, the addition of $A_{f\geq}$ to the model noticeably increased the correct classification of alveolar stops.

5.3.2 Example

Figure 5-4 shows the three lowest convex peak frequencies as a function of time, with their amplitudes, at the release of a /d/ by a female speaker. The three peak traces have been labeled F1, F2, and F3, but it is likely that the lowest peak does not follow the first vocal tract formant until at least 10ms after release.

Table 5.11 shows a series of log-ratios comparing the probability that figure 5-4 is the syllable /dah/ to the probability that it is /gah/. Each row of the table compares the probabilities of the /dah/ and /gah/ models producing a frequency or amplitude track

Difference in Log Likelihood between /dah/ and /gah/ models

Time (ms)	0	10	20	30	40
F1	-0.06	-0.02	0.26	0.94	1.56
A1	-0.02	0.35	0.12	0.62	0.59
F2	0.22	-0.39	0.05	0.06	0.11
A2	4.51	4.21	3.88	3.52	3.54
F3	-0.17	0.34	0.69	1.08	1.32
A3	0.36	0.17	0.25	0.15	0.10
$A_{f\geq}$	0.34	0.26	-0.05	-0.17	-0.40
Energy	-0.05	0.17	0.35	0.67	0.73
Total	5.13	5.10	5.54	6.86	7.54

Table 5.11: Log-likelihood scores used to decide whether the token in figure 5-4 is a /d/ or a /g/. The table shows the difference in log-likelihood between /dah/ and /gah/ models for each formant frequency model and each formant amplitude, computed independently.

equal to the corresponding convex peak frequency or amplitude shown in figure 5-4. These probabilities are calculated cumulatively, from left to right, using the forward algorithm, with each frequency and amplitude computed independent of the others.

Notice that some of the parameter tracks match the model for /d/, some match the model for /g/, and most are rather ambivalent. The $A_{f\geq}$ parameter favors the alveolar model toward the beginning of the syllable, but as the parameter $A_{f\geq}$ drops, so does the probability of an alveolar interpretation. The low values of A2 appear to be strong evidence for an alveolar interpretation: if this syllable were velar, there would be a strong peak in the burst at the frequency of F2, and A2 would therefore be much higher early in the syllable.

The total log-likelihood of /dah/, as opposed to /gah/, can be estimated by adding all of the component log-likelihoods. [1] As shown, the addition of several component log-likelihoods strongly favors the /d/ interpretation over the /g/ interpretation. The ratio of the two probabilities is exp(7.54) $\approx$ 2000.

5.3.3 Training from a large speech corpus

The parameters of the modified formant model described above were trained using all voiced stop releases in the 18 right contexts described in table 5.4 in the TRAIN subdirectory of TIMIT, and excluding releases which cross a word boundary: a total of 3141 stops. Since manual transcription of this many stop releases was judged to be impractical, the model parameters were trained on the basis of automatic formant measurements, generated the Entropic formant tracker, and smoothed using the rule-based algorithm described in chapter 3. These formant frequencies were matched to peaks in the corresponding DFT spectra, and missing information was estimated simultaneously with the model parameters using an iterative expectation-maximization algorithm.

[1] The sum of independent component log-likelihoods is not quite equal to the total log-likelihood which would be computed by the complete stochastic formant model. In the complete model, the amplitude and frequency of a formant are not independent, so it is not possible to separate the cumulative log-likelihoods as shown in table 5.11.

Speech data (merging similar contexts)

The mean, variance, and temporal correlation of each modeled variable were trained independently for each of the context cells defined in section 5.2.1. As noted in that section, many of the context cells in the TRAIN subdirectory of TIMIT are too sparse to allow reliable training of parameters.

In order to improve parameter training, context cells were collected into groups which make sense for the training of each parameter. For example, the model of each formant frequency was trained using data from all consonants with the same place of articulation, regardless of the consonant manner. Formant amplitude, spectral energy, and high frequency amplitude models were trained using all tokens with a given phonetic transcription, regardless of the speaker gender.

Even after collapsing irrelevant dimensions as described above, some of the training cells were still too sparse for reliable model training. These sparse cells were supplemented by tokens with similar right contexts, as discussed in section 5.2.1.

Matching LPC and DFT peaks

The training database included 3141 voiced stop releases, extracted from the TRAIN subdirectory of TIMIT. Since manual transcription of this many stop releases was judged to be impractical, the model parameters were trained on the basis of automatic formant measurements.

Formants were first measured, for each training token, using the Entropic LPC-based formant tracker, and the resulting LPC roots were smoothed using the rule-based algorithm developed in chapter 3.

Since the HMM output model looks only at the DFT spectrum, all formants proposed by the LPC-based algorithm were matched with peaks in the DFT spectrum, and formant means and variances were computed using the peak frequencies and amplitudes of the corresponding DFT peaks. Each formant was matched one-to-one with the closest DFT peak, provided that the closest peak was within 400 Hertz.

Any formant sample at any point in time which could not be matched with a DFT peak was marked as missing. Approximately 12% of the formant samples in voiced stops, 20% of the formant samples in unvoiced stops, and 11% of the formant samples in nasal release syllables could not be matched to a DFT peak, and were marked as missing.

Estimating missing information

The contribution to the model parameters of formants missing from the DFT spectrum was estimated using an iterative expectation-maximization algorithm, due to Johnson and Wichern (1992). Like the Baum-Welch re-estimation algorithms usually used in HMM training, the algorithm used here can be viewed as a gradient maximization of the expected log-likelihood of the training data (Rabiner, 1993). Unlike the Baum-Welch algorithm, however, the algorithm used in this section treats all formants as known, except those which have no corresponding DFT peak.

In the algorithm used here, the mean, variance, and temporal correlation of a frequency or amplitude parameter are first initialized using the statistics of the known, measurable formants in the training data. The initial values of the mean and covariance are then used to predict the contribution to mean and variance of the missing formant frequencies and

Classification using an HMM formant tracker

V	C	NT	/b/	/d/	/g/
Vowels and Glides	/b/	771	91		
	/d/	632	13	74	13
	/g/	602		12	81
Retroflex	/b/	85	96		
	/d/	81	19	62	20
	/g/	60	13		80
Lateral	/b/	69	80	12	
	/d/	11	18	82	
	/g/	42	24		67

Table 5.12: Confusion matrices for voiced stops, classified using the stochastic formant model. Entries of 10% or less have been omitted.

amplitudes, and the predictions are used to update the model statistics. This process is iterated until the parameter statistics stabilize.

In general, the initial mean, variance, and temporal correlation parameters changed little during re-estimation.

5.3.4 Classification of place

The two-level HMM formant models of each syllable, trained as described above, were used to classify the place of all voiced stop releases in the 18 right contexts of table 5.4 in the TEST subdirectory of TIMIT, excluding releases which span a word boundary: a total of 1253 stops. The three places of articulation were treated as being *a priori* equally likely.

Classification in vowel and glide context

Table 5.12 describes the result of classifying voiced stop releases using the modified stochastic formant algorithm. In vowel and glide contexts, classification is 83% correct. This result is similar to the 84% correct classification achieved by Nossair and Zahorian (1991) using similar measures (automatically measured formant frequencies and amplitudes). In their study, Nossair and Zahorian used a context-independent classifier, which presumably made their classification task more difficult, but their speech data was composed of isolated monosyllables, which presumably made classification less difficult.

The stochastic formant model was intended to improve on the performance of standard formant-based classifiers, by allowing "hidden" formant frequencies to evolve even when there is no corresponding spectral peak in the output. In training on a large database, however, several simplifications were made, which tied the "hidden" formants of the stochastic formant model much more closely to observed peaks in the spectrum. Perhaps as a result of these changes, the classification performance of the modified stochastic formant model does not exceed the classification performance reported by Nossair and Zahorian on a context-independent (and therefore more difficult) task.

Dependence of classification on right context

The LDA classifier in section 5.2 performs poorly in retroflex and lateral contexts if only formant frequencies are available, but does reasonably well if measurements of the burst

V	spoken	NT	b	d	g
LB	/b/	63	92		
/aa,	/d/	29		83	
ah/	/g/	64		17	81
HF	/b/	216	94		
/iy,y,	/d/	102		76	14
ih,ux/	/g/	33		15	79
HB	/b/	18	83		11
/uw,w,	/d/	25		88	12
uh/	/g/	31			84

Table 5.13: Stochastic formant model classification of voiced stops, as a function of right context. Entries of 10% or less have been omitted.

spectrum are available. As shown in table 5.12, the stochastic formant model does reasonably well in retroflex and lateral contexts: 80% correct in retroflex context, 76% correct in lateral context. Apparently, the formant amplitudes measured by the stochastic formant model carry enough information about the burst to allow reasonably good place classification in liquid context. The pattern of errors in retroflex context is slightly different from that of the LDA classifier: alveolar stops are almost equally likely to be called labial or velar, and velar stops are more likely to be called labial than alveolar.

Voiced stop classification as a function of vowel context is shown in table 5.13. The data show most of the same patterns seen previously with the linear discriminant classifier, and with Winitz' human subjects.

5.4 Summary

This chapter has reported context-independent and context-dependent LDA classification of place using rule-based formant and burst spectral measurements, and context-dependent maximum likelihood classification using a stochastic formant model.

Linear discriminant classification of automatic formant and burst spectral measurements results in 76% correct context-independent classification (section 5.1), and 84% context-dependent classification (section 5.2). Experiments in section 5.1 suggest that, of the 24% error in the context-independent classifier, more than half is the result of measurement error, and up to one fourth is caused by use of LDA classification, which is suboptimal.

If these sources of error can be eliminated, the experiments in sections 5.1 and 5.2 suggest very speculatively that context-dependent classification using formant and burst spectral measurements may result in better than 96% correct classification of place. This speculation is supported by the work of Kewley-Port (1982), who found that formant frequencies alone were sufficient to classify place of a stop with 97% accuracy, given context information.

It is not easy to eliminate the error in automatic formant measurements. Having failed to eliminate measurement error, this chapter suggests two other methods for working around it.

First, section 5.1 suggests analyzing the effect of error using an aggregate error model. An aggregate error model might be appropriate in a phonetic study, or in a speech recognizer based on an expert system. In either of these applications, a detailed error model can be used to predict the reliability of composite statistics, such as the linear discriminant statistic analyzed in section 5.1.

Second, section 5.3 suggests the use of a stochastic formant model, which may avoid some of the effects of measurement error by refusing to explicitly measure the formant frequencies. Unfortunately, simplifications described in this chapter apparently force the "hidden" formant models to follow explicit convex peaks in the DFT spectrum, in effect forcing the stochastic formant model to track explicit estimates of the formant frequencies. The resulting 83% correct classification of the stochastic formant model is similar to the classification performance of an LDA classifier using explicit formant and burst spectral measurements.

Chapter 6

Conclusions

This thesis demonstrates, in several stages, the relationships by which articulatory variability entails acoustic variability, and by which acoustic variability is manifested in the forms of acoustic measurement error and phonetic classification error.

As predicted in chapter 1, automatic measurements of formant frequencies and burst spectral measurements have not proven optimal in phonetic classification experiments, because of the effects of measurement error. Although these measurements are not optimal for classification, however, they have proven quite useful in demonstrating the links between articulatory variability, measurement error, and classification error. The Bayesian uncertainty models considered in chapter 4, for example, are reasonably successful in predicting measurement error based on measurements of the ambiguity in the acoustic spectrum. Classification error, in turn, is shown in chapter 5 to be well predicted by the aggregate statistical error models developed in chapter 3.

Section one of this chapter discusses the classification results from chapter 5, and compares them to the results obtained in previous classification studies. Section two discusses the error modeling results, and explores possible applications. Finally, section three discusses possibilities for future development of a stochastic formant model of human speech perception.

6.1 Classification of Consonant Place

The context-dependent classification results reported in chapter 5 are, for every type of acoustic measurement classified, very similar to previously reported results using context-independent classifiers, while the context-independent results reported in chapter 5 are slightly worse. Two factors seem to contribute to the difference in results. First, most of the published studies in chapter 1 are based on recordings of isolated or stressed monosyllables, while the experiments in chapter 5 are based on speech data excised from TIMIT, which presumably contains much more prosodic and contextual variability. Second, the failure of the context-dependent classifiers in chapter 5 to surpass previously published context-independent results can perhaps be attributed to the training procedure, which may not have allowed the trained classifiers to make full use of context information.

6.1.1 Context-independent classification

Section 1.3 reports a rather remarkable fact: most previous studies of context-independent classification of consonant place agree, to within about 5%, on the amount of place information contained in various acoustic cues. According to this consensus of researchers, burst spectral measures can be used to classify place with roughly 80-85% accuracy, formant frequencies can be classified with 65-70% accuracy, a combination of formant frequencies and amplitudes can be classified with 80-85% accuracy (Nossair and Zahorian, 1991), and dynamic spectra can be classified with greater than 90% accuracy.

In this thesis, vowel context stop release tokens in the Error Modeling database (appendix A.2) are classified in several context-independent classifiers. Formant and burst spectral information measured by human judges on this database are used to classify place with 89% accuracy, depending on the classifier design. This result matches the results of Lamel (1988), who obtained approximately 90% correct classification of place using knowledge-based classification of similar measurements by a human judge. Automatic measurements of formant frequency and burst spectral information in this thesis are substantially less reliable, with only 76% correct classification.

The 89% correct classification of manual transcriptions reported in this thesis is roughly comparable to the best speaker-independent classification reported by any study using any set of measurements. We can therefore tentatively conclude that the combination of formant frequencies and burst spectral information, if measured as human judges measure it (and provided that the recording level is appropriately controlled or normalized), is a complete representation of the consonant place information available in the release waveform.

Unfortunately, errors in the automatic measurement of formant and burst information are apparently so frequent that they make classification difficult. The 76% classification rate reported in section 5.1 is no better than the classification rates reported previously using onset spectral measures alone (e.g. in Blumstein and Stevens, 1979), even if we allow a margin of error to account for the suboptimal classifier structure.

Section 5.1 demonstrates that the difference between the 89% correct classification of manual measurements and the 76% correct classification of automatic measurements can be attributed to measurement error. A listing of the measurements which are most responsible for classification error is difficult to provide in a formal experiment, but can be provided heuristically through an analysis of section 3.4. For example, it might be speculated that a classification error occurs whenever a measurement of the onset of F2 or F3, or of the front cavity resonance in any alveolar or velar token, is mistaken by more than about 300Hz. By multiplying probabilities complementary to those shown in figure 3-11, we find that 300Hz errors in these three measurements would cause classification errors in about 28% of tokens. This is slightly larger than the 24% error rate actually observed, implying that measurement redundancy helps to reduce the error rate somewhat.

6.1.2 Context-dependent classification

Context-dependent classification rates in this thesis are generally similar to the context-independent classification rates reported elsewhere. In section 5.2, formant frequencies are classified correctly 72% of the time (compare to 65-70% in context-independent studies), burst spectral measures are classified correctly 80% of the time (compare to 80-85% in context-independent studies), and a combination of both is classified correctly 84% of the time. In section 5.3, formant frequency and amplitude information is classified correctly 83% of the time, which is comparable to the 84% achieved by Nossair and Zahorian (1991) in a context-independent study.

It is not at all clear why context-dependent results in this thesis match previous context-independent results so closely, for all types of measurements. Of the studies reviewed in section 1.3, the only study which examined context-dependent classification (by Kewley-Port, 1982) found that context information improved classification using manually transcribed formant frequencies by about 30%, from 68% correct to 97% correct. In contrast, the context-dependent LDA classifier in section 5.2 of this thesis performs only 8% better than the equivalent context-independent classifier in section 5.1.

Speculatively, it seems possible that the training attempted in chapter 5 does not make effective use of context information. In that chapter, sparse context cells are supplemented by tokens with similar right contexts, so that many of the classifiers are trained using tokens from 2-4 right contexts. Perhaps this pooling of tokens during training eliminates some of the benefit of context information.

If we assume that all of the context-dependent classifiers in chapter 5 are only about 8% better than equivalent context-independent classifiers would be, the data in chapter 5 can be taken as approximate confirmation of the previously reported results. Classification of formant frequencies is roughly 60-70% correct, and the classification of either burst spectral measures, alone, or formant frequencies and amplitude, together, is 75-85% correct, depending on the variability of the speech data. Classification of formant frequencies and burst spectral measures, taken together, is also 75-85% correct.

6.2 Predicting the Effect of Measurement Error

Although this thesis is unable to eliminate measurement error, it demonstrates repeatedly that it is possible to characterize and predict measurement error using quantitative models.

Three rather different modeling methods are presented. First, production models developed in chapter 2 are used to help design the knowledge-based measurement algorithms in chapter 3. Second, aggregate error models developed in chapter 3 are used to predict the confusion matrix of a context-independent linear discriminant classifier in chapter 5. Finally, an HMM formant tracker is used in chapter 4 to develop novel real-time measurement uncertainty models.

6.2.1 Production models of the sources of error

Chapter 2 reviews and extends several production models which, in combination with empirical data from previous studies, are used to make rough quantitative predictions of the form and range of formant and front cavity resonance measurement error. The sources of error examined include random variation of a turbulent spectrum, increased bandwidth of formants during aspiration, and the presence in the spectrum of back cavity and subglottal resonances.

By modeling turbulence noise as a Gaussian random process, and assuming a uniformly distributed phase spectrum, section 2.1.4 derives a χ^2 model of the probability distribution of the time-averaged power spectrum. Based on this model, random spectral nulls are shown to be much more likely than random spectral peaks. In an example spectrum, consisting of independent 330Hz bands (measured as the average of two 6ms Hanning windows), the χ^2 model predicts that one out of every 100 measured bands contain a randomly generated spectral peak of 5.2dB or more, while the same number of measured bands contain a null of at least 11dB.

Based on a simplified model of the contribution of glottal losses to formant bandwidth, chapter 2 demonstrates that a doubling in average glottal area can lead to an increase by four in the first formant bandwidth. It is argued that such a large increase is sufficient to reduce the Q of F1 during aspiration to nearly unity, making the formant difficult to find in the spectrum.

The influence of subglottal resonances on LPC is predicted through analysis of the LPC prediction error at the frequency of a formant peak. Based on this analysis, it is predicted that LPC will usually only track a subglottal resonance if a nearby formant is hidden by formant merger, or if the amplitude of a nearby formant is less than the amplitude of the subglottal resonance. Data from a study by Ishizaka et al. (1976) are cited to show that a subglottal resonance is unlikely to have a larger amplitude than any nearby formant, unless the formant amplitude is reduced by glottal losses during aspiration, or by a random spectral null.

Finally, the transfer function amplitudes of front and back cavity resonance peaks in frication are derived. It is demonstrated that the amplitude of a high-frequency front cavity resonance is inversely proportional to the square of the resonance frequency, while the amplitude of a back cavity resonance depends primarily on the separation of the pole and zero. Based partly on empirical data and partly on the models, it is argued that the transfer function amplitude of a front cavity resonance peak is usually larger than the transfer function amplitude of a back cavity resonance peak, but that this is not always true if the front cavity resonance frequency is sufficiently high.

6.2.2 Predicting classification on the basis of aggregate error models

After developing knowledge-based algorithms for the measurement of formant frequencies and burst spectral information, chapter 3 addresses the issue of measurement error in these algorithms. Several models of the measurement error are developed. A simple additive Gaussian error model is shown to be insufficient, for two reasons. First, measurements of formant and front cavity resonance frequencies are affected by occasional nonlinear peak-picking errors, which typically show up as "outliers" which are too large to be predicted by a Gaussian distribution. Second, measurements of burst amplitude, and of the onset frequency of F1, are heteroskedastic, that is, the measurement errors are correlated with the correct value of the measurement.

Section 5.1 demonstrates the use of quantitative error models in predicting the performance of a linear discriminant classifier. A simple additive Gaussian error model is insufficient for accurate prediction of the confusion matrix, but a heteroskedastic Gaussian model predicts the confusion matrix reasonably well.

The prediction of discriminant error from measurement error in section 5.1 shows one way in which aggregate error models, of the sort developed in chapter 3, might be used in a larger knowledge-based speech recognizer. A knowledge-based recognizer built, for example, using a blackboard expert system (e.g. Carver and Lesser, 1992) usually depends on quantitative estimates of the reliability of various competing recognition hypotheses. Section 5.1 demonstrates the use of quantitative error models to predict the reliability of several binary classifiers, and of the round-robin classifier built from them; similar prediction based on quantitative error models might be useful in judging the reliability of hypotheses in a blackboard-based speech recognizer.

Simple quantitative models are currently used to describe the difference between transcribers in large acoustic phonetic studies. In their study of vowel formants, for example, Hillenbrand et al. (1995) qualified their formant frequency measurements with a table describing the mean absolute difference between the measurements of different transcribers. Sections 3.4 and 5.1 suggest that such simple models may not be sufficient for all purposes, because of error outliers or heteroskedasticity. In particular, section 3.4 suggests that studies which rely on automatic measurements of any kind should qualify their results with a carefully tested error model.

6.2.3 Real-time predictions of measurement uncertainty

Section 6.2.2 suggests that aggregate error models of the type developed in chapter 3 might be useful in predicting the reliability of competing hypotheses in a blackboard-based speech recognition system. An aggregate error model, however, only contains general information about the type of measurement being made. A blackboard expert system might get more benefit from an error model which includes information about quirks of the specific token being recognized, including the distribution of spectral peaks in each band of interest.

The stochastic formant model developed in chapter 4 generates specific uncertainty models which predict the possible errors in each and every formant measurement, as it is made. These uncertainty models are designed to pass on information about the quirks of the token under study, including the distribution of peaks in each band of interest, to any higher level recognition algorithm which might employ the model.

The stochastic formant model can also be used to classify speech sounds by itself, without being fitted into a larger speech recognizer. Section 5.3 implements a classification model in which formants are viewed as production states, which may or may not match peaks in the acoustic spectrum at any given time. This model is designed to avoid the problem of measurement error by avoiding measurements: stochastic models of the formants associated with each place of articulation are used to classify a stop release waveform, without ever explicitly measuring the formant frequencies of the waveform being classified.

The classification performance of the stochastic formant model in section 5.3 is almost exactly equal to the classification performance obtained using LDA classification of explicit formant and burst spectral measurements. There are two likely reasons for the similarity in performance between these two algorithms. First, the stochastic formant model is modified, in chapter 5, to accept a list of convex peak frequencies and amplitudes as input in lieu of the entire DFT spectrum. Speculatively, it may be that the reason the stochastic formant model is unable to beat the performance of an LDA classifier is that the input to the stochastic formant model and the input to the LDA classifier contain similar lists of spectral peak frequencies and amplitudes. Second, the stochastic formant model in chapter 5 is trained on data generated by the knowledge-based formant measurement algorithm, so it is entirely possible that the formant models trained in chapter 5 are only able to learn spectral distinctions which are also already captured in the knowledge-based formant frequency representation.

Apparently, we do not yet have enough information about the stochastic formant model to judge its usefulness in speech classification. The next section will describe future work which might help to evaluate the usefulness of stochastic formants in a speech classifier.

6.3 Future Work: Stochastic Formant Models of Perception

Perceptual studies indicate that listeners can hear cues offered via formant frequencies (e.g. Delattre, Liberman, and Cooper, 1955). On the basis of these perceptual studies, phoneticians have occasionally argued that human speech perception includes some kind of low-level formant tracking module, which passes information about the formant frequencies of a signal up to higher-level classification modules.

Unfortunately, nobody has ever been able to build a formant tracking algorithm which is sufficiently free of errors to be a plausible model of a low-level perceptual process. When a formant tracker misses a formant, it usually proposes a pattern of formants which is quite different from the correct formant pattern. When human listeners make vowel identification mistakes, on the other hand, the formant patterns of the proposed vowel are usually similar to the pattern of the correct vowel (Huang, 1991). Formant trackers always make mistakes, but human listeners almost never make the kinds of mistakes that they would make if perception depended on a formant tracker. Therefore, most phoneticians today do not believe that perception depends on a low-level formant tracker.

The stochastic formant model provides a completely new model of the way formant tracking might be used in speech perception. In the stochastic formant model, there is no formant tracker, and hence the algorithm is not affected by formant tracking errors (at least as the model is presented in chapter 4; the convex peak representation in chapter 5 can be viewed as a kind of formant tracker). Instead of explicit formant measurements, each phoneme model is composed of implicit formant *models*, each of which is tasked with explaining the distribution of energy in its own frequency band. If a formant peak is missing from the spectrum, the other formants are not shifted up or down to compensate. Instead, the phoneme model adds together the uncertainty of the missing formant, the certainty of those formants which are clearly measured, and the information available from non-formant components (e.g. energy) to compute a total recognition score. Thus the stochastic formant model is designed to make misclassification errors of one or two distinctive features, as do human listeners, rather than making arbitrary complete phoneme errors, as do formant tracking algorithms.

The stochastic formant model has not been aggressively portrayed as a model of perception in this thesis for two reasons. First, the models proposed in this thesis use a linear frequency scale. Before the stochastic formant model can be credibly related to human perception, it will have to be re-designed to accept an input similar to the input accepted by the human perceptual system. The frequency scale should be warped to fit an auditory criterion; some kinds of masking and gain control may also prove to be useful. A nonlinear frequency scale may result in loss of formant frequency resolution at high frequencies; some experimentation will be needed to determine whether this loss of resolution is a help or a hindrance to correct phoneme classification.

Second, the stochastic formant model has not been aggressively portrayed as a model of perception because the prototype classifier tested in chapter 5 did not work very well, compared to existing speech recognition systems which use no explicit formant or sub-band information of any kind. It was speculated, in section 6.2.3, that the algorithm's poor performance in chapter 5 is probably due to the reduced dimensionality of the input, and to the dependence of the training procedure on possibly erroneous formant contours produced by the knowledge-based formant smoothing algorithm. These speculations should be tested: the model should be trained directly on DFT spectra, using some kind of parameter re-estimation algorithm, and tested in a phoneme classification task using a complete observation space.

Finally, if the model is to be convincing, it should be tested against a control: a more established speech classification algorithm, as much like the stochastic formant model as possible, but with no knowledge of the formant structure of speech. If the observation space consists of 50ms from a mel-frequency spectrum, for example, a suitable control algorithm would be a frame-based continuous-density model of the mel-frequency spectrum or cepstrum, with variation over time in the density means, perhaps something like the nonstationary HMM proposed by Deng et al. (1994).

Appendix A

Speech Data

A.1 KB Train: Training tokens for knowledge engineering

The KB Train database is composed of 180 consonant releases selected from the TRAIN subdirectory of TIMIT. These 180 consonant releases include 20 releases of each consonant, 10 each spoken by male and female speakers, in arbitrary context. This is the only database in this thesis which contains consonant releases which span a word boundary.

The tabulated lists below give the TIMIT filename of each sentence, the consonant identity, and the release time. Release times were re-transcribed during acoustic analysis, so the release times given may not correspond exactly to thc release times transcribed in TIMIT. Ten manual acoustic measurements are listed. For each token, the first three formants 20ms and 50ms after stop release are shown. For stops, the low-frequency and high-frequency burst amplitudes (lfa and hfa) and number of peaks (np) are shown. Finally, for alveolar and velar bursts, the burst front cavity resonance (fcr) is listed.

b releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fbcg1/si982	1436		5	-15	-20	607	733	1340	1508	2806	2764
fbcg1/si982	838		3	-9	-15	398	411	2576	2468	2995	3044
fcag0/si2133	1170		1	-25	-29	378	517	2050	2189	2428	2587
fceg0/si618	611		6	-21	-39	678	884	1830	1789	2797	2920
fclt0/si808	1210		4	-14	-17	375	365	2219	2300	2625	2625
fclt0/si808	2517		10	-12	-9	375	376	2219	2250	2781	2812
fdml0/sx339	1195		7	-25	-26	469	443	2062	2031	2906	2844
fgmb0/si515	1928		3	-3	-10	535	377	1933	2094	2674	2953
fhlm0/si2190	1323		4	-19	-27	500	529	1469	2062	2781	3125
mapv0/si1293	1448		2	-2	-9	656	667	1125	1250	2312	2219
marw0/sx349	3268		1	-3	-15	531	596	1344	1437	2187	2219
mbbr0/si1685	1115		1	-23	-28	503	503	1298	1340	2136	2367
mbom0/si1014	1517		4	-19	-28	594	681	1031	1000	2406	2344
mbom0/si1014	3085		7	-17	-19	500	568	1562	1562	2125	2250
mbom0/si1014	134		4	-10	-12	344	425	1812	1875	2437	2469
mbsb0/si723	2576		8	-7	-11	469	562	1100	1312	2437	2375
mbsb0/si723	3088		2	-22	-18	375	512	1281	1200	2250	2200
mcae0/si2077	2838		6	-30	-35	500	472	1219	1219	2312	2281
mctm0/si1980	1082		6	-13	-14	562	646	1250	1281	2500	2344
mctm0/si1980	2319		5	-1	-7	562	532	1281	1437	2406	2875

d releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fbmj0/si1776	754	3281	2	-10	-8	406	424	2281	1719	2875	2719
fbmj0/si1776	2918	3250	3	0	3	344	495	1906	1500	2625	2656
fcmg0/si1872	1596	6187	8	-4	-1	250	376	2687	2844	3062	3187
fdfb0/si1318	1164	5531	3	-15	-12	500	636	1969	1969	3125	3062
fdkn0/si1202	3162	3687	5	-5	-4	375	496	1937	2000	2719	2750
fdml0/sx339	145	4523	4	-10	-16	250	330	2062	2156	2812	2875
fdnc0/si1278	3091	4219	5	-5	4	375	518	2281	2531	3031	3250
mapv0/si1293	4488	3875	4	1	2	531	578	1469	1625	2437	2406
marw0/sx349	4476	3687	3	-6	-5	344	455	1500	1562	2375	2469
mbbr0/si1685	886	4126	6	-37	-40	432	658	1995	1851	2612	2550
mbbr0/si1685	2211	3687	5	-7	-5	469	541	1812	1844	2437	2500
mcss0/si688	329	4084	6	-31	-16	461	461	1759	1822	2681	2702
mdbp0/si528	1328	3435	5	-5	1	411	482	1707	1571	2509	2471
mdbp0/si528	3411	4250	3	-8	-1	281	352	2031	1969	2562	2594
madd0/sx178	1411	2062	2	10	-17	312	463	1406	1469	1937	1850
mapv0/si1293	966	2719	2	8	-4	344	384	1625	1500	2250	1500

g releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fcag0/si2133	1411	2586	2	5	-4	438	546	2110	1939	2627	2626
fdkn0/si1202	4080	2687	1	8	-9	375	554	2750	2125	3125	2600
fdkn0/si1202	1844	2906	4	8	-1	375	439	2219	2156	2969	2656
fdml0/sx339	1421	2000	1	1	-10	281	399	1906	1562	2781	2781
fdtd0/si931	1121	3219	1	0	11	500	601	2650	2687	3125	3125
fecd0/sx338	1239	2031	3	12	-2	400	469	2344	2400	2800	2687
fhlm0/si2190	2481	1406	1	6	-10	406	508	1469	1469	2000	2187
fjen0/si2307	1382	1719	2	12	6	531	757	1781	1719	2469	2594
fjen0/si2307	3024	1656	1	13	-8	375	616	1844	2000	2100	2000
mcae0/si2077	4503	1250	2	-3	-20	437	568	1281	1281	2125	2500
mcae0/si2077	2141	1344	1	-5	-20	312	405	1406	1344	2062	2062
mcxm0/si1351	1064	1594	2	-21	-37	312	347	1656	1594	1656	1800
mdlc0/si1395	1762	937	1	6	-38	344	444	1031	1156	1875	1937
mdlc0/si1395	2170	1187	1	-13	-32	344	472	1219	1250	1600	1719
mdlc0/si1395	4547	875	1	1	-33	375	404	1031	1062	1750	1875
mdrd0/si752	2589	2531	3	7	-2	406	517	2250	2031	2250	2531
mhjb0/si2277	1192	2094	1	-3	-15	344	450	2219	1937	2500	2375

p releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fbcg1/si982	1642		2	-2	-6	531	645	1000	781	2550	2700
fbcg1/si982	2561		5	-12	-17	500	437	1719	1937	2281	3300
fbmj0/si1776	3373		9	-17	-21	406	455	1625	1250	2000	2500
fcrz0/si793	434		6	1	-1	281	596	1375	1281	2625	2625
fear0/si1252	1949		6	-10	-13	600	724	1344	1250	2156	2300
fexm0/si1731	704		9	-17	-19	625	555	1437	1969	2312	2656
fjkl0/si932	2520		2	-2	-18	594	614	1475	1531	2000	1800
fjkl0/si932	437		1	2	-11	312	545	1562	1844	2219	2281
fjlg0/si1506	594		2	4	-1	500	478	1406	1406	2125	1406
fjlr0/sx241	973		9	-29	-29	250	351	1844	2031	3062	3156
fjrb0/si1302	2943		8	-17	-16	187	473	1906	1875	2562	2437
marw0/sx349	242		4	-5	-23	312	372	1125	1125	1437	1600
mcss0/si688	1456		4	1	-2	594	672	1156	1000	2406	2531
mcss0/si688	3600		5	2	-2	562	578	1094	1437	2250	2344
mcss0/si688	1842		4	-4	-10	656	637	1687	1625	2969	2906
mcss0/si688	750		8	-6	-11	350	391	1542	1748	1933	1995
mdbp0/si528	434		6	-13	-20	531	600	1125	1125	2500	2650
mdhs0/si2160	922		5	-24	-25	400	450	1625	1406	2625	2500
mdrd0/si752	144		8	-20	-21	250	354	1094	1200	1531	1406
mdwd0/si1890	3174		7	-15	-17	500	574	1219	1281	2562	2687

t releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fapb0/si2323	4939	4781	5	0	0	469	601	1906	1781	2844	2875
fapb0/si2323	2079	5125	10	-12	-9	687	779	1750	2062	2594	2625
fapb0/si2323	1262	4281	7	-5	-1	375	412	2100	2250	3156	2906
fbcg1/si982	2704	4906	7	0	2	875	955	2094	2000	3406	3400
fbcg1/si982	133	5562	10	-7	1	531	710	2200	2250	3500	3125
fcag0/si2133	1073	4020	3	-18	-16	378	404	1811	1374	2368	2242
fceg0/si618	517	4156	6	3	6	312	284	1781	1406	2687	2344
fceg0/si618	1568	2974	4	9	-5	500	460	1219	800	2437	2312
fceg0/si618	294	4219	6	3	8	625	800	1800	1406	3000	2844
fceg0/si618	1325	4531	7	-1	4	281	474	2250	2687	3219	3150
fceg0/si618	2126	5437	8	-1	6	250	459	2100	2125	2937	2750
fclt0/si808	1735	5257	5	-1	-8	437	468	1937	1812	2594	2406
fcmg0/si1872	1833	5156	6	-2	-2	700	735	2625	2719	3312	3050
fdas1/si2091	372	4500	1	-20	-9	800	710	2400	2094	3400	3031
fdas1/si2091	629	3219	4	-12	-8	531	627	1900	2062	2900	2937
fdfb0/si1318	1478	5437	4	-35	-21	375	431	1969	2281	3031	2844
fdfb0/si1318	139	4594	8	-7	0	187	432	1781	1100	3062	2875
fdkn0/si1202	2163	4687	8	-1	-2	250	450	2000	2200	2800	2906
fdkn0/si1202	534	5031	7	1	6	281	527	1900	1900	2875	2875
fdnc0/si1278	2225	4531	2	-9	4	500	515	2200	2156	3219	3219
madd0/sx178	3165	4031	4	0	3	344	417	1719	1719	2437	2281
madd0/sx178	1641	4500	8	-3	3	375	426	1800	1781	2594	2500
mapv0/si1293	2972	3625	5	-6	-4	437	619	1562	1562	2312	2437
mbsb0/si723	773	4562	4	4	8	500	550	1850	1250	2950	2812
mcew0/sx182	1548	3156	6	0	3	312	368	1700	1656	2437	2437
mctm0/si1980	757	3727	3	10	1	514	535	1296	1069	3455	3002
mdbp0/si528	2279	4126	4	8	11	628	482	1780	1738	2660	2408
fbcg1/si982	2352	2187	3	7	-6	531	597	1937	1719	1937	2187
fclt0/si808	1412	3031	2	1	10	687	720	1656	1406	2250	2050
fclt0/si808	3122	3281	2	-6	7	437	440	2281	1625	3406	3350
fcrz0/si793	1897	2366	1	5	-12	531	605	2250	1906	2500	2062
fdkn0/si1202	2812	3250	6	1	2	437	492	1800	1844	2469	2100
fdnc0/si1278	4921	3187	2	-2	4	450	573	1812	1750	1812	1750
mdbp0/si528	2278	3969	5	-1	2	312	404	1900	1750	2687	2406

k releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fapb0/si2323	4155	2875	6	-9	-14	312	595	2187	2187	2906	2812
fbcg1/si982	654	2312	1	3	-11	375	464	2450	2469	2812	3094
fbjl0/si2182	677	2062	3	-1	-10	562	559	1844	1687	1844	2062
fbmj0/si1776	3205	2969	3	2	-5	250	331	3594	2600	3594	3812
fcmg0/si1872	303	2812	7	-22	-24	500	470	3031	2700	3531	3437
fdfb0/si1318	2527	1594	1	2	-10	469	591	1375	1469	1800	2094
fdjh0/si2195	1911	1469	1	-1	-12	437	480	1812	2094	1812	2094
fdkn0/si1202	903	2000	1	0	-15	469	558	2031	1937	2031	2375
fdml0/sx339	637	2687	1	-13	-25	281	685	2500	2500	2844	2844
fdnc0/si1278	676	2094	1	0	-13	1500	1422	1937	1687	3250	3062
fdnc0/si1278	971	1781	1	1	-15	450	496	1594	1687	1594	1687
madd0/sx178	2863	2750	7	-5	-7	781	670	1812	1469	2344	2156
madd0/sx178	556	1719	1	10	-5	406	486	1750	1750	2500	2300
mapv0/si1293	436	1594	1	13	-1	469	536	1625	1344	1625	1600
mapv0/si1293	3492	1625	1	9	-3	437	644	1656	1344	2150	2000
marw0/sx349	4314	1812	2	7	-4	344	293	1700	1656	2375	2344
marw0/sx349	963	2062	4	13	9	281	473	1719	1406	2344	2250
marw0/sx349	569	1656	1	-8	-21	344	422	1687	1437	2312	2062
mbbr0/si1685	1726	1469	1	1	-11	531	612	1875	1562	2375	2187
mbbr0/si1685	138	1656	1	2	-15	719	909	1781	1375	2187	2125
mbef0/si651	2613	1500	1	0	-15	500	496	1344	1031	2150	1937
mbef0/si651	886	1406	1	-16	-31	844	553	1000	850	2187	2300
mbom0/si1014	2743	1156	1	-13	-32	875	758	1100	1094	1906	2406
mcss0/si688	1268	2812	3	-2	-6	312	366	2500	2031	3156	3125
mctm0/si1980	3108	1625	1	-1	-13	719	676	1625	1406	2375	2031
mdbp0/si528	4097	1312	1	-1	-20	719	706	1469	1437	2125	2937

m releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fapb0/si2323	3413					474	441	2001	2024	2817	2926
fbjl0/si2182	1740					777	774	1194	1353	3090	3136
fbmj0/si1776	3547					484	486	1754	1860	2737	2871
fclt0/si808	672					779	766	1879	2143	2591	2742
fcmm0/si453	1137					431	409	1610	1661	2766	2795
fdjh0/si2195	2095					662	630	2086	2351	2779	3088
fdkn0/si1202	1593					616	484	1674	1698	2732	2766
fdxw0/si881	2210					413	416	773	726	3193	3168
feeh0/si1112	3015					319	351	2676	2457	3167	3028
fltm0/si1700	1819					538	557	766	1014	3639	3034
madd0/sx178	1818					527	518	1731	1870	2602	2606
mapv0/si1293	4020					570	571	1486	1637	2192	2252
marw0/sx349	3732					582	566	1234	1269	1789	1827
mbbr0/si1685	279					692	712	1076	1211	2321	2175
mbef0/si651	485					514	556	724	864	2072	1975
mbom0/si1014	1058					570	598	1621	1661	2549	2521
mcae0/si2077	1285					393	387	1904	1964	2617	2648
mdcd0/sx155	1189					522	544	1089	1262	2264	2331
mdlc0/si1395	503					543	545	1310	1448	2468	2551
mdrd0/si752	1684					576	548	1942	2070	2706	2724

n releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fapb0/si2323	1435					916	887	2344	2455	2912	3060
fbcg1/si982	395					565	560	2199	2135	3227	3148
fbjl0/si2182	183					756	821	2006	1923	2864	2841
fblv0/si1688	2830					555	573	1952	1813	2781	2679
fcag0/si2133	685					768	812	1642	1466	2932	2918
fceg0/si618	975					471	683	1050	1155	3473	3299
fcke0/sx121	1334					428	405	1999	1979	2792	2781
fcmg0/si1872	2180					836	796	1977	1969	3206	3238
fcrz0/si793	880					684	693	1551	1167	2923	2992
fpaf0/si2314	1636					480	552	2422	2179	3029	2580
mafm0/sx309	1439					609	625	1352	1358	2557	2592
makb0/sx206	743					636	627	1114	798	2031	2238
mapv0/si1293	1985					512	538	1520	1560	2613	2542
mbgt0/si1841	1975					355	347	2022	2078	2641	2633
mcae0/si2077	4897					603	592	1321	1287	2528	2543
mcss0/si688	2199					525	539	1307	1294	2808	2683
mdbp0/si528	2579					383	378	2282	2191	2833	2866
mdhs0/si2160	402					620	633	1795	1725	2703	2506
mfxs0/si2304	175					687	768	1893	1739	2658	2766
mmea0/si758	1075					354	369	962	721	2769	2430

ng releases

Filename	Time	fcr	np	lfa	hfa	F1		F2		F3	
fblv0/si1688	699					625	696	2276	1842	2916	2925
fcke0/sx121	709					419	436	2226	1959	2645	2609
fdxw0/si881	2034					477	507	1612	1168	2883	3094
fetb0/sx248	547					577	283	2478	2003	3216	3013
fgmb0/si515	1633					767	775	2400	2201	3169	2910
fjdm2/sx142	1171					688	747	1447	1450	2929	2810
fjlg0/si1506	1144					764	736	2438	2455	3048	3061
flma0/si1873	2378					603	484	2028	1369	2957	2544
ftaj0/si699	657					617	672	1780	1670	2806	2683
maeb0/si2250	1291					434	496	835	845	2137	2148
mafm0/sx309	1281					504	434	1824	1575	2163	2261
makb0/sx206	635					643	651	1793	1537	2269	2269
mbgt0/si1841	1836					378	400	2076	1961	2631	2490
mdcd0/sx155	2265					502	500	1750	1670	2580	2469
mesj0/si997	2594					324	314	1758	1826	2340	2437
mjeb1/si837	1185					300	330	2074	2011	2869	2686
mjwt0/si1291	2863					641	609	963	1069	2325	2413
mkah0/si1528	1604					426	437	2190	1925	2514	2525
mmsm0/si1106	785					509	504	2073	1567	2696	2698
mmwb0/si2249	2197					435	505	2220	1950	2673	2738

A.2 Error Modeling: Data used to develop error models

The Error Modeling corpus is used for three purposes in this thesis. First, automatic rule-based measurements of this corpus are compared to the manual transcriptions given below, and the difference is used to train models of the measurement error in section 3.4. Second, this database is used in section 4.4 to test the *a posteriori* uncertainty distributions generated by the stochastic formant model. Finally, manual and automatic measurements of the vowel context stops in this database are classified, using LDA and other classifiers, in sections 1.1 and 5.1.

The tabulated lists below give the TIMIT filename of each sentence (the si/sx specification has been omitted to save space), as well as the release time, right context (V), and ten acoustic measurements. Release times of nasals are as indicated in the TIMIT transcription; release times of stops have been re-transcribed. Most tokens are extracted from the TEST subdirectory of TIMIT, but a few combinations of context and consonant which did not exist in the TEST subdirectory were extracted from the TRAIN directory. None of the consonant releases in this database span a word boundary (according to the TIMIT transcription), but some span word-internal syllable boundaries.

Acoustic measurements listed are a combination of measurements by two human judges; if both judges transcribed a given token, their measurements have been averaged. Listed measurements include all three formants 10ms and 50ms after release, the burst front cavity resonance (fcr), burst low-frequency and high-frequency amplitudes (lfa and hfa), and burst diffuseness (df). Blank measurements indicate a formant which did not exist in the spectrum, or for some other reason could not be measured.

b releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fjsj0/404	139	aa		5	37	36	440	775	1026	1089	2534	2492
fjas0/50	1633	ae		5	54	57	508	672	1607	1787	2344	2557
fisb0/1579	2394	ah		4	57	58	465	626		1778		2748
fdrd1/1566	714	ax		5	42	29	404	505	1252	1212	2828	3353
flbw0/1849	587	axr		5	34	32	485	485	1535	1737	2020	2222
fcmh1/1493	1587	eh		5	53	48	525	788	1879	2101	2768	2828
fjwb0/992	3302	el		3	52	50	343	424	808	990	3091	2869
fjsj0/134	502	er		4	49	40	566	525	1111	1616	1737	1939
fjsj0/404	1528	ih		5	51	52	303	465	1717	2000	2465	2606
fasw0/380	2324	ix		4	44	38	546	525	1717	2101	2606	2606
fcmh0/14	137	iy		5	64	61	404	404	2101	2444	2647	2869
fdrw0/1423	1740	l		3	41	49		465	1111	909	3050	3131
fgwr0/2208	2750	r		5	49	51	525	647	1354	1515	2062	2141
fhew0/223	469	uh		5	26	28	323	505	1374	1636	3172	3131
fpkt0/908	1550	uw		5	56	57	444	525	1313	1333	2444	2667
fgmd0/143	1059	ux		5	60	59	242	384	1919	2182	2707	2970
fnip0/408	1039	w		3	61	53	525	586	990	1374	2869	2808
fjlm0/53	1329	y		3	41	39	404	404	1960	2202	2788	2869
mwjg0/404	141	aa		4	46	34	502	595	888	953	2233	2159
mljb0/50	2451	ae		4	43	41	424	748	1172	1414	2082	2263
mpam1/576	135	ah		4	38	38	454	625	878	1241	2317	2307
mnjm0/230	1474	ax		4	39	38	492	554	1128	1205	2000	2046
mjvw0/1733	3089	axr		5	51	45	426	461	1082	1257	1801	1774
mmdb1/2255	1703	eh		5	47	45	451	554	1482	1533	2621	2656
mrjm4/319	1406	el		5	40	48	446	405	830	810	2440	2554
mbpm0/317	510	er		5	37	31	436	518	1257	1360	1722	1764
mmwh0/279	1089	ih		5	43	43	420	512	1338	1082	2312	2820
mhpg0/460	719	ix		4	42	38	456	554	1533	1605	2241	2615
mctt0/298	1781	iy		5	49	47	343	317	1954	2117	2482	2713
mrjm4/49	2223	l		4	32	40	404	369	1051	1149	2444	2482
mreb0/385	1392	r		5	35	40	424	505	1010	1172	1475	1535
mplb0/2024	1086	uh		5	37	38	451	456	836	810	2543	2610
mrcz0/911	1368	uw		5	46	47	375	323	892	856	2407	2430
mrjm4/859	806	ux		4	43	47	343	358	1596	1790	1923	2410
mroa0/1970	1273	w		5	39	33	707	748	1212	1535		
mljb0/140	2881	y		4	42	38	283	297	1556	1743	1919	2354

d releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fdrd1/1544	1368	aa	4586	5	61	63	485	808	1919	1616	3091	2808
fjsa0/749	2959	ae	4525	4	68	67	546	647	2061	2121	2990	3010
fmld0/822	412	ah	2162	3	59	55	566	748	1737	1636	2335	2303
fcmh1/2123	1080	ax	4081	5	46	48	424	950	2364	1960	3253	3152
fjcs0/1309	3878	axr	3434	5	58	60	303	424	1495	1556	2788	1899
fjmg0/101	1477	eh	4323	5	61	68	404	707	2283	2404	3131	3111
fjlm0/413	1321	el	3264	4	60	62	497	605	1924	1189	2984	2832
fpas0/224	1021	er	3232	3	49	61	444	505	1778	1758	2566	2081
flnh0/224	529	ih	4162	5	56	67		364	2303	2424	3192	3293
fjas0/2030	643	ix	4242	5	58	61	283	404	2040	1980		2586
flbw0/1219	1438	iy	4505	4	63	71	283	444	2283	2101	2929	2808
fnmr0/319	1178	l	4869	5	70	74	364	485	1475	1212	2929	2707
ftlh0/199	688	r	3292	5	56	62	384	586	1576	1596	2667	2121
fsjg0/940	136	uh	3494	4	74	77	283	404	2020	1879	2667	2303
futb0/34	598	uw	3939	3	70	79		424	1879	1657	3232	2929
futb0/214	661	ux	3757	3	59	76	404	404	1980	1677	3010	2889
fjsa0/119	1433	w	4990	5	52	55	263	384	1374	1071	3212	3192
fecd0/788	4199	y	4586	2	44	56	222	404	2586	2606	3455	3010
mroa0/1970	2376	aa	2995	2	48	57		533	1495	1199	2485	2502
mpwm0/407	534	ae	3815	4	63	68	539	580	1810	1908	2646	2564
mpgl0/1099	404	ah	5717	5	38	51	407	467	1461	1374	2750	2574
msfh1/100	603	ax	3337	3	52	60	476	467	1909	1600	2876	2564
mjes0/2014	3162	axr	2662	2	51	37	444	415	1657	1513	2283	1846
mgjf0/101	1466	eh	3805	4	62	64	407	543	1732	1569	2662	2275
mrml0/251	1878	el	2133	2	56	34		405	990	733	2505	2202
mbwm0/224	1015	er	2800	2	60	47	436	430	1559	1472	2245	1846
mrjr0/2313	4026	ih	4467	5	52	55	407	415	1974	1913	2785	2723
mfgk0/214	2890	ix	4748	5	42	49	333	446	1780	1815	2744	2759
mmwh0/1301	139	iy	4677	4	58	66	297	317	2061	2241	2682	2943
mjfc0/1033	1235	l	3010	4	46	48	262	236	1102	1118	2492	2523
mgjf0/191	291	r	3394	4	61	62	657	405	1507	1297	1939	1610
mjfc0/43	332	uh	3969	5	43	45	359	369	1211	1312	2870	2292
mcsh0/1549	1229	uw	3297	2	56	62	323	405	1754	1544	2543	2512
mjes0/214	633	ux	4163	4	46	59		359	1733	1610	2621	2405
mrms1/857	1924	w	2803	1	50	35	424	430	828	810	1455	2087
mdwa0/185	259	y	3466	4	53	61	317	333	1903	1887	2641	2420

g releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fsem0/1828	4286	aa	2378	3	48	52	605	865	2356	1989	2919	3157
fsem0/28	2258	ae	3157	4	66	68	541	670	2465	2249	3330	3178
fjwb0/365	1946	ah	2909	4	73	70	566	808	1939	1657	2950	3071
flkd0/289	863	ax	1596	1	76	63	485	727	1616	1313	2929	3253
fcau0/227	1715	axr	1596	2	68	65	465	546	1697	1616	2404	1798
fpkt0/98	142	eh	2808	1	69	63	364	727	2465	2040	2808	2687
felc0/1386	4142	el	1293	1	62	61	364	546	1374	1131	2667	2849
fcmh0/104	1712	er	1717	2	48	52	364	546	1778	1616	2283	2040
ftlh0/379	465	ih	2909	1	62	49		444	2748	2586	3030	3152
fcau0/317	1133	ix	2505	3	66	68		424	2343	2162	2707	2828
fsem0/208	2399	iy	3438	3	68	64		432	2984	2789	3568	3524
futb0/214	1878	l	1091	1	65	53		343		1333		2626
fmml0/410	140	r	1576	1	69	60		546	1657	1959	2364	2323
fjsj0/224	334	uh	1737	1	65	62	303	546	1879	1919	2465	2626
fetb0/338	1124	uw	1172	4	58	61		465	1859	1616	3010	3030
fblv0/338	1368	ux	1939	2	58	59	303	384	1980	2020	2626	2444
frng0/95	731	w	1050	1	59	51		505	1051	889		2525
frng0/725	2933	y	3212	1	62	55		586		2343	3333	3232
mctt0/28	554	aa	1328	2	67	57	562	554	1425	1144	2121	2107
mjth0/396	685	ae	2517	4	47	48	359	425	2518	2430	3140	2830
majc0/205	2471	ah	1790	1	58	56	404	543	1818	1400	2162	2133
mrjm4/139	443	ax	1092	2	52	53	497	456	1061	979	2036	2200
mdwk0/2170	2436	axr	1877	1	60	50		430	1899	1523	2182	1984
mglb0/904	3060	eh	2733	3	54	50	323	492	2162	1877	2586	2528
mrcs0/143	1427	el	1092	2	38	32	282	405	1134	918	1980	2518
mjfc0/223	1519	er	1497	2	50	42	339	405	1404	1164	1635	1507
mpgl0/379	501	ih	2327	1	57	54	375	405	1968	1800	2364	2364
mjes0/754	4422	ix	2485	1	66	56	343	424	2061	1798	2667	2525
mmjr0/208	2933	iy	2882	2	51	46	276	359	2222	2236	2960	2943
mmdb1/995	1699	l	1118	1	61	53		394	1221	1272	2759	2753
mdbb0/1195	1356	r	1374	1	68	46		343	1313	1253	1859	1616
mpgl0/469	2168	uh	1701	3	52	56		440	1657	1733	2263	2395
mjhi0/338	1128	uw	1266	1	63	55	333	523	1225	1123	2141	2436
mrms1/857	2337	ux	1559	1	59	45		405	1616	1558		2343
mjdm1/95	278	w	1071	1	62	51		482	922	836		2027
mtab0/42	1538	y	2472	2	58	61	303	317	2404	1953	2820	2355

p releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fmah0/119	2009	aa		5	45	47	869	909	1576	1677	2869	3091
fdhc0/929	1664	ae		5	54	54	687	667	2182	2263	2667	3111
fjwb0/635	4122	ah		5	39	41	687	889	1111	1293	2667	2849
fcmh0/1454	3445	ax		5	56	51		303	1677	1838	2808	2889
fpas0/404	1015	axr		4	53	49		444		1091	1313	1535
fdhc0/29	1270	eh		5	52	56		637	2020	1910	2727	2746
fjsj0/314	413	el		4	47	48		444	1313	990	2626	2909
fnlp0/138	2855	er		5	57	57		647	1515	1556	2465	2384
fdms0/138	1227	ih		5	27	30	485	525	2040	1737	3051	3172
ftlh0/1390	1672	ix		4	72	69		505	1737	1737	2869	3030
fdms0/1848	2259	iy		5	56	57		343	2444	2849	3030	3252
fisb0/229	1385	l		2	51	61			1151	1071	2889	2970
fmml0/140	1031	r		3	58	46			1354	1313	2061	1576
fjre0/216	926	uh		5	54	54		485	1152	1495	2424	2707
fdrw0/383	1324	uw		5	41	42		424	1455	1616	3131	3030
fpls0/330	1237	ux		4	36	42	404	465	1535	1818	2586	2525
fasw0/200	1862	w		5	44	49	525	606	748	1010	2788	2808
fhes0/299	1557	y		2	55	49	384	465	2101	1879	2950	2768
mglb0/364	3599	aa		5	46	48	719	672	1118	1062	2312	2241
mpam0/379	137	ae		4	56	59	875	656	1759	1722	2458	2502
mgwt0/1539	1021	ah		5	38	42		875	1051	1067	2416	2436
mtls0/290	2447	ax		5	45	49		647	1455	1576		2101
mrjr0/102	3235	axr		5	24	30		487	1434	1333	2283	1515
mcmb0/278	836	eh		4	45	42	782	549	1569	1718	2154	2282
mdlf0/53	669	el		4	46	44	444	502	946	759	2697	2765
mrtk0/373	1335	er		5	45	44		518	1215	1236	1838	1635
mrgg0/29	923	ih		4	58	56	456	492	1712	1677	2405	2251
mrgg0/29	1608	ix		5	50	52	528	369	1441	1533	2329	2395
mabw0/314	306	iy		3	61	64	594	343	1944	2087	2779	2898
mrjm4/229	964	l		4	43	48	750	518	1128	850	3111	2292
mkdr0/643	1560	r		5	42	46			1143	1275	1466	1639
mjth0/216	991	uh		4	52	50		477	1326	1520	3071	2933
mbns0/50	2622	uw		5	42	45		395	1082	918	2444	2446
mdrm0/383	1252	ux		5	44	45		395	1410	1513	2456	2389
mjtc0/200	1475	w		5	30	28		626	657	990	2505	2272
mahh0/124	2149	y		4	59	59				2141	2584	2723

t releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fjlm0/323	2371	aa	4849	5	51	69			1778	1737	3212	3192
fjmg0/281	1666	ae	5090	5	57	78			2546	2546	3414	3354
felc0/2016	2774	ah	3980	4	57	68		727	1859	1636	2849	2485
fcft0/278	135	ax	3131	4	64	61			1758	1737	3131	2667
fslb1/284	2100	axr	2727	4	72	70		424	1939	1737	2687	2303
fjre0/126	792	eh	4990	5	57	76				2061	3111	2990
fgmd0/683	418	el	5414	2	56	72			1798	1394	3091	3192
flas0/1026	1193	er	4081	3	45	68		546		1899	2667	2283
fjre0/1116	548	ih	4869	5	47	79		626	1879	1596	3051	2788
fmah0/29	657	ix	4626	5	60	63	323		2525		3192	
fadg0/109	2259	iy	4757	4	58	74		454	2119	2616	3027	3157
fcdr1/376	1858	l	3737	3	40	41		343		1333	2586	2444
fjmg0/281	797	r	4061	3	58	78					2828	2141
fasw0/1550	710	uh	3697	2	68	81			2384	2121	3091	3091
fcmr0/205	2734	uw	3737	4	62	71			2162	2061	3152	2889
futb0/1204	2319	ux	4748	4	57	78			2081	1879	3030	2950
flbw0/49	1006	w	3556	4	66	74			1758	1556	2566	2384
ftaj0/699	790	y	2849	4	53	45			2081	2040	2869	2727
mdwa0/95	2142	aa	3759	3	65	64	469	625	1708	1318	2333	2487
mpab0/1103	1993	ae	3918	3	60	59		652	1815	1795	2744	2636
mbns0/590	520	ah	5030	4	42	58		672	1569	1441	2641	2564
mpam1/396	920	ax	4533	5	52	55	782	467	1877	1232	2543	2415
mres0/137	142	axr	2811	2	56	50	750	407	1702	1708	2444	2230
mjfc0/2293	2748	eh	4465	5	58	64		505	1616	1535	2404	2323
mlnt0/1574	1973	el	4728	5	54	64		750	1737	1220	2465	2286
mjtc0/830	748	er	4236	5	56	65	719	562	1859	1800	2723	2424
mpcs0/1359	1162	ih	4528	4	53	68		456	1770	1728	2795	2713
mjmp0/365	1379	ix	4081	5	67	73		566	1758	1636	2343	2444
mmwh0/279	1354	iy	4169	5	53	62		343	1759	2061	2713	2573
mwvw0/396	1058	l	3974	3	34	47		844	1650	1703	2973	2667
mjes0/394	1757	r	2123	4	55	45		415	1576	1246	1835	1528
mcem0/2028	301	uh	4141	5	46	68			1616	1434		2444
mdbb0/295	3795	uw	3626	4	59	68		750	1859	1682	2587	2590
majc0/295	3101	ux	4476	4	54	63	782	750	1764	1728	2467	2600
mrjm3/368	889	y	3164	4	61	57		407	2202	1988	2648	2506

k releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fslb1/891	1677	aa	2465	4	69	66			2323	2081	2889	2929
faks0/1573	808	ae	2263	4	72	72		889		2040	2444	2707
fgwr0/2208	672	ah	3151	2	58	54		546	2343	2162	2808	2788
fkms0/50	1001	ax	2767	2	59	56		606	2081	1677	2849	2768
fgwr0/948	3573	axr	1838	1	75	64		465	1899	1535	2242	1899
fmcm0/550	562	eh	2909	1	61	51			2485	2626	3151	3252
fadg0/109	1646	el	1313	1	67	58		707	1333	1192	2586	3131
ftlh0/1009	738	er	2081	2	72	68			2162	1939	2768	2323
fgmd0/143	1478	ih	2748	1	73	70	303	465	2323	2020	2768	2505
faks0/133	1517	ix	1758	3	73	74		444	2020	1737	2687	2707
fisb0/229	703	iy	2808	1	75	66		505	2606	2626	3111	3051
flas0/138	251	l	1576	1	70	55		808	1414	1273	3152	2970
flnh0/941	2975	r	1253	1	74	61			1374	1495		2020
fasw0/290	1793	uh	2061	2	68	64		444	1960	1657	3252	3010
fsbk0/349	687	uw	1212	1	73	63		485	1333	1232	2970	2950
fhes0/1109	3016	w	1576	1	65	59		303	1434	1071	2788	
fjmg0/191	1683	y	3535	4	59	55			2929	2687	3596	3657
mglb0/94	926	aa	1595	2	59	50	469	605	1657	1420	2246	2307
mrjs0/94	1237	ae	2966	4	61	61	719	646	1937	1948	2868	2677
mrjo0/734	2996	ah	1657	2	55	56		562	1806	1420	2586	2236
mtmr0/133	942	ax	1636	1	56	55	721	467	1615	1338	2271	2071
mrpc0/933	1341	axr	1369	2	55	53		497	1523	1384	1697	1692
mmjr0/208	1981	eh	2518	2	58	56		497		1774	2513	2471
mjmp0/95	2058	el	1287	1	67	62		440	1111	870		2121
mjbr0/101	1586	er	1585	2	55	39	500	502	1482	1287	1818	1538
mmdm2/102	2376	ih	2733	3	58	54		395	2242	1851	2563	2369
mplb0/44	393	ix	2586	4	59	60		384		1677	2586	2465
mjln0/99	2453	iy	3409	2	60	60		236	2444	2236	3273	3138
mcmj0/374	509	l	1215	1	59	46		525	1118	1212	1778	
msfh1/640	3548	r	1738	1	63	50		594	1657	1497		2475
mctw0/2003	1223	uh	1333	1	69	53		500	1302	1241		2220
mjdm1/455	1309	uw	1701	4	55	56	469	389	1533	1041	2182	2451
mjvw0/113	450	ux	2569	2	53	50	907	385	2020	1738	2584	2354
mrjm3/98	572	w	1066	1	57	44		354	1102	953		2311
mdab0/409	2243	y	2667	4	53	58		327	1778		2404	2503

m releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
felc0/756	1808	aa					681	673	1128	1241	3219	3308
fcmh1/413	2215	ae					849	970	2020	1980	2969	3071
fjmg0/551	547	ah					736	788	1795	1838	2849	2788
fedw0/274	2512	ax					666		1495	1717	2889	3092
fnlp0/138	1492	axr					546	485	1394	1717	2202	2222
fmld0/205	2427	eh					626	707	1717	1838	2485	2869
fhlm0/120	3578	el					497	497	908	887	2530	2595
fgwr0/138	1023	er					626	606	1495	1515	1919	2020
fhes0/299	1083	ih					606	465	2404	2020	3071	2990
fcft0/188	1272	ix					566	485	1798	2040	2929	3111
fcrh0/1718	1367	iy					485	485	2444	2687	3111	3354
fjem0/364	1618	l					546	627	849	995		3049
fsms1/64	2007	r					566	525	1859	2586	2384	2950
fblv0/1058	2105	uh					505	546	1152	1354	3010	3030
fksr0/397	1483	uw					485	465	1434	1495	2909	2950
fram1/730	3429	ux					364	384	2485	2424	2889	2929
fhes0/209	1446	w					541	735	908	1060	2357	2314
fisb0/319	3693	y					384	505	2465	2283	3394	2626
mdrm0/23	1758	aa					525	656	785	902	1887	1887
mdsc0/1038	1136	ae					554	625	1400	1472	2230	2333
mjvw0/113	219	ah					595	641	1005	1236	2605	2621
mdrb0/94	1245	ax					631	219	1139	1407	2420	2282
mhpg0/280	1050	axr					625	636	1005	1066	1420	1380
mnjm0/950	4784	eh					683	728	1615	1579	2282	2375
mgwt0/1539	1282	el					462	467	723	820	2677	2672
mjrf0/371	2300	er					564	543	1189	1312	1820	1702
mcem0/768	3916	ih					456	467	1780	1836	2256	2297
mkjl0/1100	2467	ix					564		1507		2046	
mnls0/1483	325	iy					405	333	1515	1960	2087	2329
mrjs0/364	862	l					477	525	1031	1487	2286	2261
mkcl0/1721	642	r					528	611	800	1011	1472	1621
mdhs0/180	270	uh					672	646	1031	1225	2091	
mfrm0/345	1629	uw					440	420	1425	1410	2109	2128
mctw0/23	434	ux					292	317	1924	1620	2364	2379
mrjo0/1364	919	w					446	508	1046	1236	1938	2061
mjvw0/23	598	y					313	297	1948	1818	2456	2451

n releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
flbw0/1219	2825	aa					828	849	1596	1374	2727	2849
fawf0/10	1698	ae					566	566	2323	2525	2889	2950
fjsa0/389	392	ah					647		1636		2505	
fhew0/223	1647	ax					606		1919		3354	
fjwb0/365	1622	axr					707	707	1899	1778	2202	2222
fcmr0/25	1632	eh					727	647	2323	2343	3010	2566
flbw0/1219	2240	el					626	788	1495	1172	2687	2586
fjsj0/314	1162	er					465	586	1697	1434	1980	1778
futb0/214	444	ih					566	546	2263	2343	2849	2828
flas0/408	1760	ix					667	748	1960	1899	3091	3051
fgmd0/413	905	iy					444	404	2222	2101	3091	2748
flnh0/404	1251	l					606	505	1475	2263	3596	3394
fnlp0/678	992	r					485	465	1455	1434	2081	2121
flkd0/894	190	uw					546	626	2303	2202	3273	2950
fasw0/920	792	ux					485	525	2404	2202	3111	3152
fawf0/1630	1106	w					444	546	828	990	3071	2849
flnh0/134	1002	y					707	525	2465	1980	3374	2626
mjjg0/373	710	aa					687	749	1462	1399	2518	2573
mpdf0/912	2092	ae					564	677	1472	1585	2590	2492
mnjm0/320	877	ah					724	564	1338	1462	2038	1971
mrgg0/389	1365	ax					503	369	1256	1092	2182	1533
mslb0/383	825	axr					611	528	1743	1677	2389	2153
mthc0/205	940	eh					590	600	1451	1538	2605	2533
mbdg0/833	997	el					621	611	943	903	2005	2073
mjmp0/5	757	er					621	590	1702	1580	2194	2011
mjvw0/203	915	ih					410	385	1636	1620	2543	2539
mcem0/2028	3157	ix					505	465	1636	1758	2606	2444
mrws1/500	2432	iy					415	343	1968	2036	2467	2579
majc0/25	289	l					543	525	1108	1596	2312	2525
mbpm0/137	1125	r					436	487	1400	1302	1615	1595
mmdh0/2286	2493	uh					595	610	1384	1462	2759	2708
mccs0/839	2261	uw					426	365	1948	1893	2585	2497
mcmj0/1094	662	ux					405	405	2026	1758	2389	2220
mlnt0/12	825	w					364	222	808	748	2283	2364
mtdt0/1994	2807	y					263	303	2020	2000	2667	2303

ng releases

Filename	Time	V	fcr	df	lfa	hfa	F1		F2		F3	
fgmd0/323	1587	aa					788	768	2121	2182	2849	2970
fclt0/268	1567	ae					670	778	2249	2119	2746	2832
fpad0/176	459	ah					808	889	1657	1737	2909	3010
fgmb0/515	967	ax					444	606	2343	2162	2889	3091
fskp0/1728	799	eh					586	626	2141	1859	2727	2687
fdrw0/1283	402	ih					546	748	2626	2283	3192	3152
fmaf0/2089	1261	ix					485		2444		2748	
flas0/48	1694	iy					525	707	2546	2566	2788	2788
ftlh0/379	1190	l					606	808	1253	1434	2748	3313
fssb0/362	788	r					586	727	1071	1535	1899	2020
fsxa0/1108	2844	w					505	485	707	808	2828	2808
fram1/730	2275	y					505	606	2485	2020	3293	2525
mrcs0/593	1987	aa					482	646	1754	1307	2543	2682
mpam1/576	1577	ae					754	713	1867	1641	2379	2359
mkch0/2008	565	ah					666	713	1415	1277	2138	2333
mbjk0/2128	4124	ax					693	303	1277	1313	2246	2141
mpam1/1029	1071	axr					502	482	1216	882	2369	2564
mdab0/2299	2128	eh					420	687	1994	1661	2297	2497
mrai0/432	1957	el					467	502	1031	995	2482	2687
mmdb1/2255	560	er					677	652	1322	1256	2538	1677
mnls0/1610	2420	ih					359	389	2149	1990	2415	2399
majc0/295	2917	ix					525	579	2000	1770	2323	2456
mcem0/48	1892	iy					456	497	2272	2077	2452	2425
mbpm0/47	683	l					533	523	1179	1812	2512	2590
mjln0/369	1480	r					349	369	1495	1071	1984	2046
mrrk0/1918	1709	w					451	430	963	929	1856	1897
mkdr0/13	512	y					404	444	1778	1596	2222	1879

A.3 SFM Train: Training data for the HMM formant tracker

In chapter 4 of this thesis, the stochastic formant model is trained as a formant tracker using manual transcriptions of a small 36-token training set (9 consonants × 2 genders × 2 right contexts, /aa/ and /ah/). All of the 36 training tokens are taken from the TRAIN subdirectory of TIMIT, and there is no overlap between this data set and the KB Train and Error Modeling data sets. None of the consonant releases in this database span a word boundary (according to the TIMIT transcription), but some span word-internal syllable boundaries.

The list below shows the TIMIT filename, release time, speaker gender (f/m), consonant (C), and right context (V) of each of the 36 training tokens, together with formant measurements 10ms and 50ms after consonant release.

Filename	Time	f/m	C	V	F1		F2		F3	
fjlr0/sx421	1833	f	b	aa	687	812	1173	1236	3062	2937
fscn0/si1886	2267	f	b	ah	406	687	1375	1375	2687	2687
mprt0/sx400	1883	m	b	aa	482	687	1047	1068	2500	2500
mmaa0/si845	231	m	b	ah	500	562	1000	1437	2750	2687
fkdw0/si1891	812	f	d	aa	500	812	1687	1437	2687	2500
flac0/sx451	2419	f	d	ah	687	687	1875	1687	2932	2687
mrws0/sx292	1064	m	d	aa	312	625	1500	1187	2375	2437
mbbr0/sx245	138	m	d	ah	-256	500	1750	1687	2375	2437
fcaj0/si1479	334	f	g	aa	-256	687	2125	1937	2492	2937
ftaj0/sx159	240	f	g	ah	437	562	2000	1562	2562	2187
mrre0/si952	3042	m	g	aa	-256	562	1937	1375	2304	1937
mwsb0/sx366	1348	m	g	ah	-256	562	1500	1312	2062	1937
fmkf0/sx366	1618	f	p	aa	625	1125	1812	1750	2937	2875
fmem0/si2007	728	f	p	ah	875	875	1625	1501	2500	2500
mdhl0/si809	1272	m	p	aa	-256	750	1312	1000	2241	2062
mtab0/sx132	747	m	p	ah	437	562	875	1125	1937	2312
fpad0/sx446	184	f	t	aa	750	875	1937	1173	3187	3187
fcyl0/si667	4395	f	t	ah	565	687	1750	1625	3125	2937
mppc0/sx332	138	m	t	aa	437	937	1812	1312	3000	2437
mlel0/sx166	2727	m	t	ah	-256	500	1625	1625	3062	2687
fgdp0/sx178	2906	f	k	aa	562	937	1812	1687	3562	3250
fsls0/sx156	2058	f	k	ah	-256	937	1937	1687	2687	2812
mtwh1/si882	2002	m	k	aa	625	1005	1437	1250	2187	2187
mclk0/sx40	1053	m	k	ah	482	562	1906	1562	2312	2437
fsls0/sx426	381	f	m	aa	650	937	838	1213	3125	3187
fcmg0/sx162	1772	f	m	ah	500	625	1250	1250	3125	3000
mbgt0/si1341	167	m	m	aa	461	699	687	843	2250	2125
mtrc0/si1623	427	m	m	ah	625	625	1131	1562	2312	2562
fpjf0/sx326	1254	f	n	aa	625	812	1750	1500	3375	3375
fgmb0/si1145	1102	f	n	ah	625	812	1687	1562	2937	2750
mpeb0/si1860	263	m	n	aa	461	625	1500	1187	2375	2187
mkjo0/si887	303	m	n	ah	500	625	1375	1187	2562	2750
fjdm2/sx142	1175	f	ng	aa	875	1131	1437	1375	2937	2812
fmpg0/si2232	1086	f	ng	ah	687	812	1875	1437	3187	3187
mrjm0/sx58	2987	m	ng	aa	625	625	1750	1375	2178	2062
mrcw0/si2001	1138	m	ng	ah	482	562	1562	1312	2312	2312

Bibliography

[Atal and Hanauer, 1971] Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2):637–655.

[Blumstein and Stevens, 1979] Blumstein, S. E. and Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Am.*, 66(4):1001–1017.

[Bush and Kopec, 1987] Bush, M. A. and Kopec, G. E. (1987). Network-based connected digit recognition. *Trans. Acoust., Speech, and Sig. Proc.*, ASSP-35(10):1401–1413.

[Carver and Lesser, 1992] Carver, N. and Lesser, V. (1992). Blackboard systems for knowledge-based signal understanding. In Oppenheim, A. V. and Nawab, S. H., editors, *Symbolic and Knowledge-Based Signal Processing*, pages 205–250. Prentice-Hall, Englewood Cliffs, NJ.

[Chen, 1991] Chen, M. Y. (1991). Acoustic parameters of nasal vowels of the hearing impaired. Master's thesis, MIT, Cambridge, MA.

[Chomsky and Halle, 1968] Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* Harper and Row, New York, NY.

[Chun, 1996] Chun, R. (1996). A hierarchical feature representation for phonetic classification. Master's thesis, MIT, Cambridge, MA.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Trans. ASSP*, ASSP-28(4):357–366.

[Delattre et al., 1955] Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am.*, 27(4):769–773.

[Deng et al., 1994] Deng, L., Aksmanovic, M., and Wu, C. J. (1994). Speech recognition using hidden markov models with polynomial regression functions as nonstationary states. *Trans. Speech and Audio Proc.*, 2(4):507–520.

[Drake, 1988] Drake, A. W. (1988). *Fundamentals of Applied Probability.* McGraw-Hill, New York, NY.

[Entropic Speech, Inc., 1993] Entropic Speech, Inc. (1993). *Entropics Signal Processing System.* Washington, DC.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production.* Mouton and Co., The Hague.

[Fant et al., 1972] Fant, G., Ishizaka, K., Lindqvist, J., and Sundberg, J. (1972). Subglottal formants. Speech Trans. Lab. Q. Prog. Stat. Rep. 1, Royal Institute of Technology, Stockholm.

[Fant et al., 1986] Fant, G., Liljencrants, J., and Lin, Q. (1986). A four-parameter model of glottal flow. Speech Trans. Lab. Q. Prog. Stat. Rep. 4, Royal Institute of Technology, Stockholm.

[Flanagan, 1972] Flanagan, J. L. (1972). *Speech Analysis, Synthesis, and Perception.* Springer, New York, 2nd edition.

[Fujimura, 1962] Fujimura, O. (1962). Analysis of nasal consonants. *J. Acoust. Soc. Am.*, 34(12):1865–1875.

[Gordon, 1969] Gordon, C. G. (1969). Spoiler-generated flow noise. II. Results. *J. Acoust. Soc. Am.*, 45(1):214–223.

[Halle et al., 1957] Halle, M., Hughes, G. W., and Radley, J.-P. A. (1957). Acoustic properties of stop consonants. *J. Acoust. Soc. Am.*, 29(1):107–116.

[Hanson, 1995] Hanson, H. M. (1995). *Glottal characteristics of female speakers.* PhD thesis, Harvard University, Cambridge, MA.

[Hanson and Stevens, 1995] Hanson, H. M. and Stevens, K. N. (1995). Sub-glottal resonances in female speakers and their effect on vowel spectra. In *Int. Conf. Phonetic Sciences*, pages 182–185, Stockholm, Sweden.

[Hillenbrand et al., 1995] Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.*, 97(5):3099–3111.

[Huang, 1991] Huang, C. (1991). *An Acoustic and Perceptual Study of Vowel Formant Trajectories in American English.* PhD thesis, MIT, Cambridge, MA.

[Ishizaka et al., 1976] Ishizaka, K., Matsudaira, M., and Kaneko, T. (1976). Input acoustic-impedance measurement of the subglottal system. *J. Acoust. Soc. Am.*, 60(1):190–197.

[Johnson, 1994] Johnson, M. (1994). Automatic context-sensitive measurement of the acoustic correlates of distinctive features at landmarks. In *Proc. ICSLP*, pages 1639–1643, Yokohama.

[Johnson and Wichern, 1992] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Englewood Cliffs, NJ, third edition.

[Kennedy, 1992] Kennedy, P. (1992). *A Guide to Econometrics.* MIT Press, Cambridge, MA.

[Kewley-Port, 1982] Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *J. Acoust. Soc. Am.*, 72(2):379–389.

[Kewley-Port, 1983] Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 73(1):322–335.

[Klatt, 1977] Klatt, D. H. (1977). Review of the ARPA speech understanding project. *J. Acoust. Soc. Am.*, 62(6):1345–1366.

[Klatt and Klatt, 1990] Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857.

[Kopec, 1986] Kopec, G. E. (1986). Formant tracking using hidden markov models and vector quantization. *Trans. Acoust. Speech and Sig. Proc.*, 34(4):709–729.

[Lamel, 1988] Lamel, L. F. (1988). *Formalizing Knowledge used in Spectrogram Reading: Acoustic and Perceptual Evidence from Stops.* PhD thesis, MIT, Cambridge, MA.

[Lehiste and Peterson, 1961] Lehiste, I. and Peterson, G. E. (1961). Transitions, glides, and diphthongs. *J. Acoust. Soc. Am.*, 33(3):268–277.

[Lin, 1990] Lin, Q. (1990). *Speech Production Theory and Articulatory Speech Synthesis.* PhD thesis, Royal Institute of Technology (KTH), Stockholm.

[Maeda, 1982] Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3-4):199–229.

[Manuel and Stevens, 1995] Manuel, S. Y. and Stevens, K. N. (1995). Formant transitions: teasing apart consonant and vowel contributions. In *Int. Conf. Phonetic Sciences*, Stockholm, Sweden.

[Massey, 1994] Massey, N. S. (1994). Transients at stop-consonant releases. Master's thesis, MIT, Cambridge, MA.

[Nossair and Zahorian, 1991] Nossair, Z. B. and Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *J. Acoust. Soc. Am.*, 89(6):2978–2991.

[Ohala, 1989] Ohala, J. J. (1989). Sound change is drawn from a pool of synchronic variation. In Breivik, L. E. and Jahr, E. H., editors, *Language Change: Contributions to the Study of Its Causes*, pages 173–198. Mouton de Gruyter, New York.

[Papoulis, 1984] Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, NY.

[Pastel, 1987] Pastel, L. M. P. (1987). Turbulent noise sources in vocal tract models. EE thesis, MIT, Cambridge, MA.

[Peterson and Barney, 1952] Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of vowels. *J. Acoust. Soc. Am.*, 24(2):175–184.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Prentice-Hall Inc., New Jersey.

[Shadle, 1985] Shadle, C. H. (1985). *The Acoustics of Fricative Consonants.* PhD thesis, MIT, Cambridge, MA.

[Stevens,] Stevens, K. N. Acoustic phonetics. In Preparation.

[Stevens, 1971] Stevens, K. N. (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *J. Acoust. Soc. Am.*, 50(4):1180–1192.

[Stevens, 1996] Stevens, K. N. (1996). Articulatory-acoustic-auditory relationships. In Hardcastle, W. and Laver, J., editors, *Handbook of Phonetic Sciences.*

[Sussman et al., 1991] Sussman, H. M., McCaffrey, H. A., and Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am.*, 90(3):1309–1325.

[Talkin, 1987] Talkin, D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs. *J. Acoust. Soc. Am.*, 82(S1):S55.

[Titze, 1994] Titze, I. R. (1994). *Principles of Voice Production.* Prentice Hall, Englewood Cliffs, NJ.

[Waibel et al., 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *Trans. Acoust. Speech and Sig. Proc.*, 37(3):328–339.

[Winitz et al., 1971] Winitz, H., Scheib, M. E., and Reeds, J. A. (1971). Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech. *J. Acoust. Soc. Am.*, 51(4):1309–1317.

[Zue et al., 1990] Zue, V., Seneff, S., and Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9:351–356.

[Zwillinger, 1996] Zwillinger, D., editor (1996). *CRC Standard Mathematical Tables.* Chemical Rubber Co., Boca Raton, Florida.