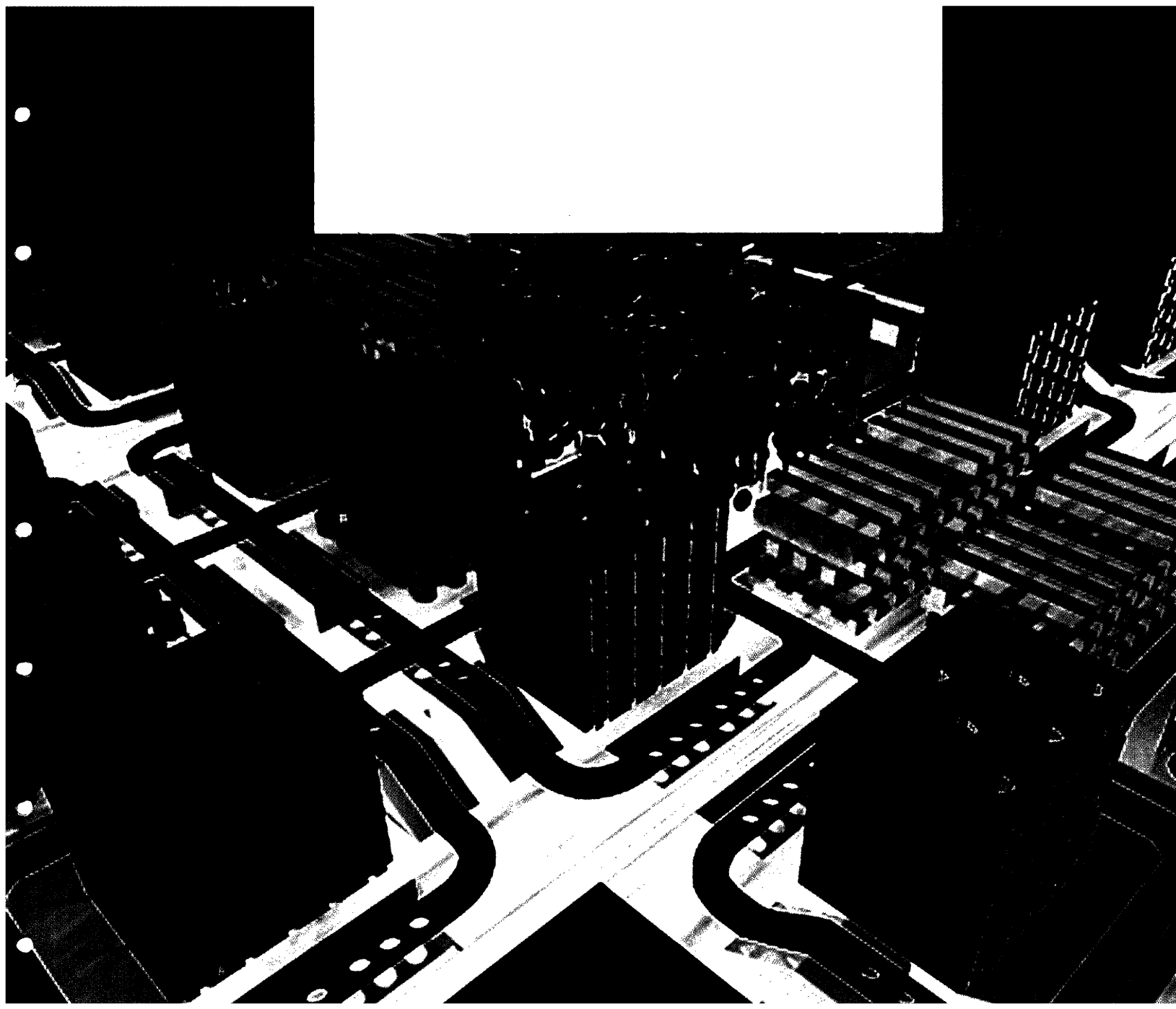


MASSACHUSETTS INSTITUTE OF TECHNOLOGY
The RESEARCH LABORATORY *of* ELECTRONICS

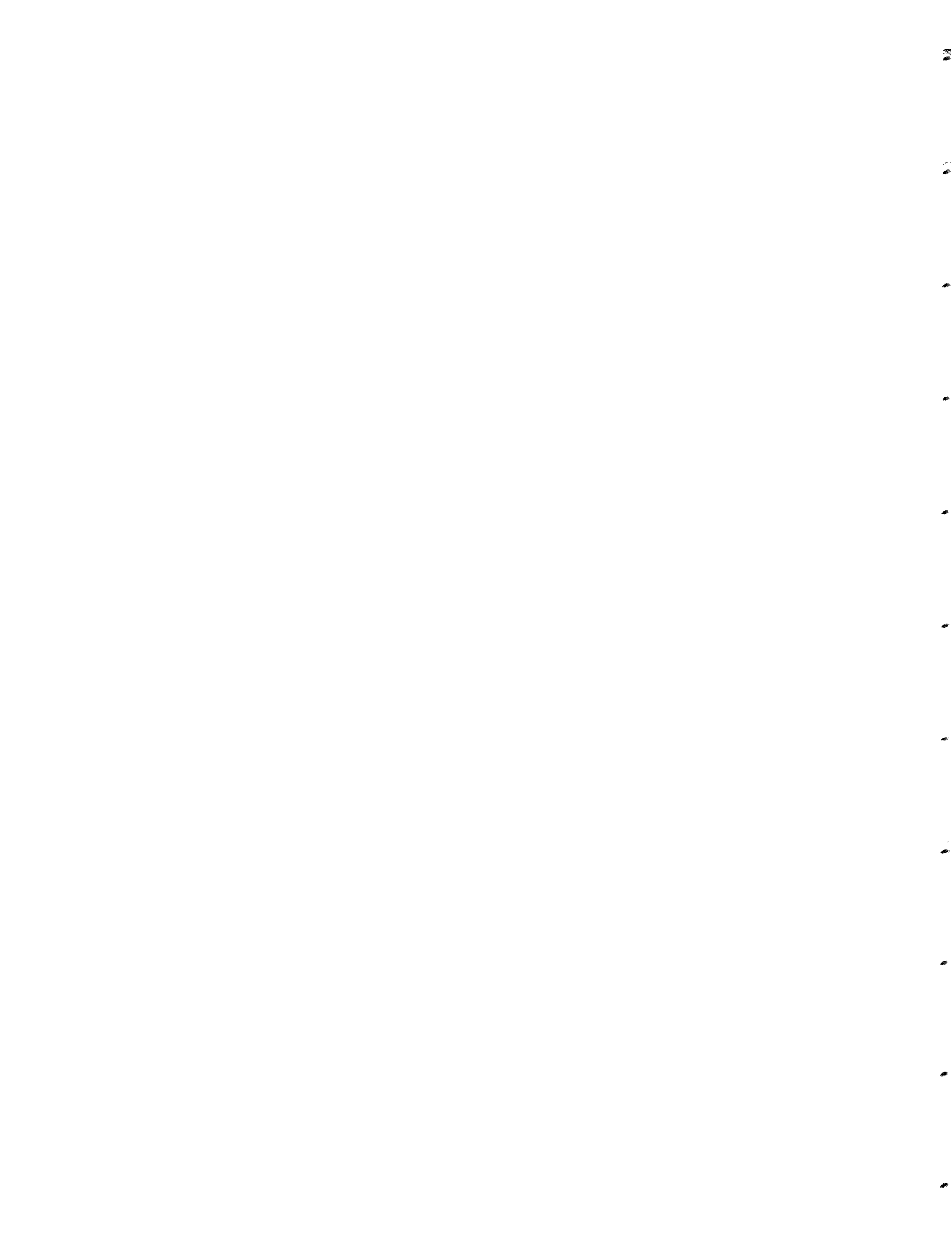


**Automatic Syllable Detection for
Vowel Landmarks**

By: Andrew Wilson Howitt

RLE Technical Report No. 642

July 2000

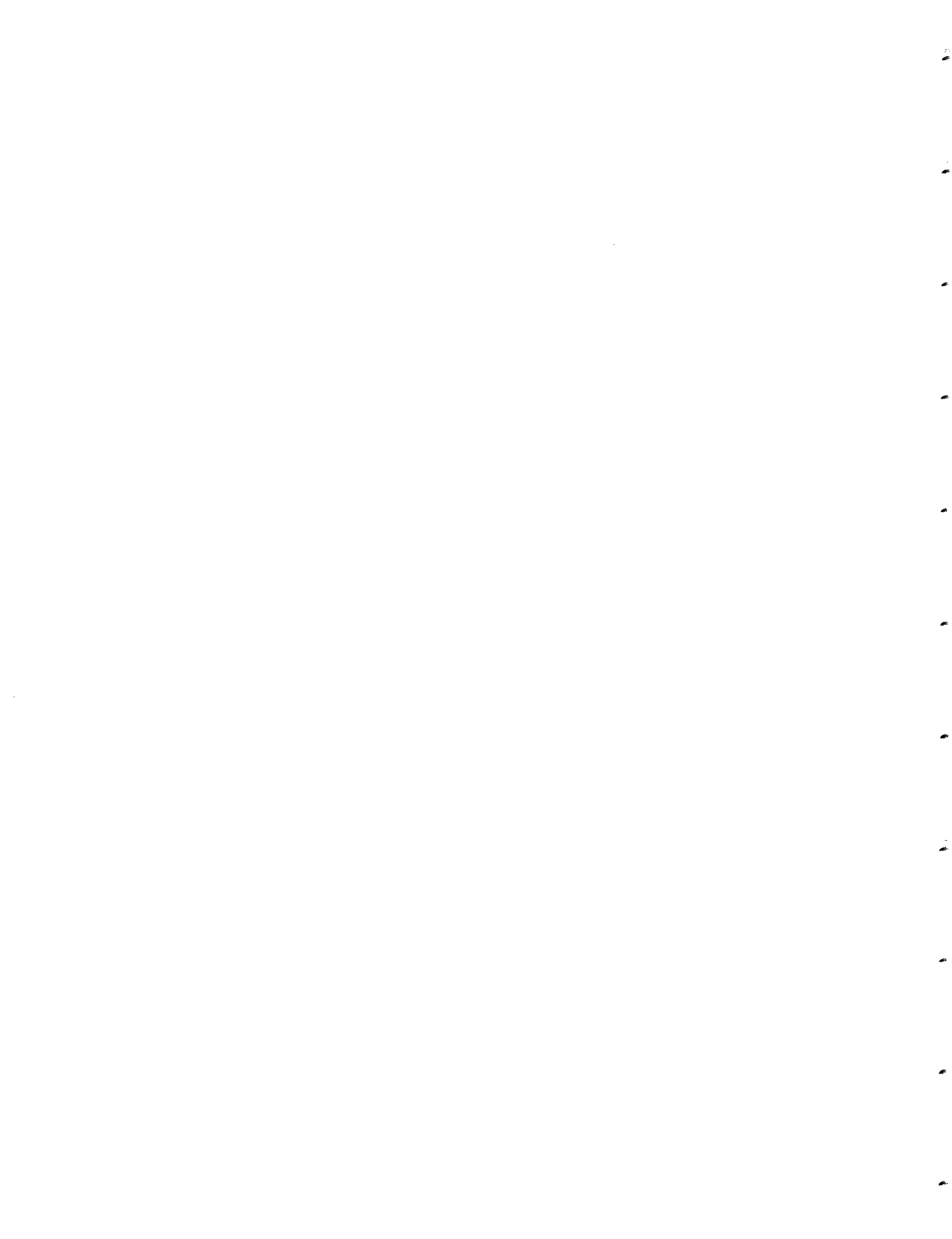


**Automatic Syllable Detection for
Vowel Landmarks**

By: Andrew Wilson Howitt

RLE Technical Report No. 642

July 2000



Automatic Syllable Detection for Vowel Landmarks

by

Andrew Wilson Howitt

B.A. Mathematics, Stonehill College (1982)
B.S.E.E., University of Notre Dame du Lac (1983)
M.S. EECS, Massachusetts Institute of Technology (1987)

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

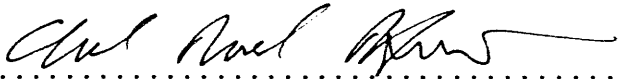
Doctor of Science

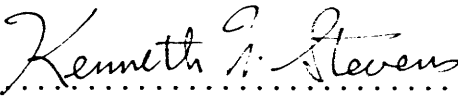
at the

Massachusetts Institute of Technology

July 2000

©Massachusetts Institute of Technology 2000. All rights reserved.

Author 
Department of Electrical Engineering and Computer Science
25 July 2000

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

Accepted by
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Automatic Syllable Detection for Vowel Landmarks

by

Andrew Wilson Howitt

Submitted to the Department of Electrical Engineering and Computer Science
on 25 July 2000 in partial fulfillment of the requirements for the degree of
Doctor of Science

Abstract

Lexical Access From Features (LAFF) is a proposed knowledge-based speech recognition system which uses landmarks to guide the search for distinctive features. The first stage in LAFF must find Vowel landmarks. This task is similar to automatic detection of syllable nuclei (ASD).

This thesis adapts and extends ASD algorithms for Vowel landmark detection. In addition to existing work on ASD, the acoustic theory of speech production was used to predict characteristics of vowels, and studies were done on a speech database to test the predictions. The resulting data guided the development of an improved Vowel landmark detector (VLD).

Studies of the TIMIT database showed that about 94% of vowels have a peak of energy in the F1 region, and that about 89% of vowels have a peak in F1 frequency. Energy and frequency peaks were fairly highly correlated, with both peaks tending to appear before the midpoint of the vowel duration (as labeled), and frequency peaks tending to appear before energy peaks. Landmark based vowel classification was not found to be sensitive to the precise location of the landmark. Energy in a fixed frequency band (300 to 900 Hz) was found to be as good for finding landmarks as the energy at F1, enabling a simple design for a VLD without the complexity of formant tracking.

The VLD was based on a peak picking technique, using a recursive convex hull algorithm. Three acoustic cues (peak-to-dip depth, duration, and level) were combined using a multi-layer perceptron with two hidden units. The perceptron was trained by matching landmarks to syllabic nuclei derived from the TIMIT aligned phonetic transcription. Pairs of abutting vowels were allowed to match either one or two landmarks without penalty. The perceptron was trained first by back propagation using mean squared error, and then by gradient descent using error rate. The final VLD's error rate was about 12%, with about 3.5% insertions and 8.5% deletions, which compares favorably to the 6% of vowels without peaks. Most errors occurred in predictable circumstances, such as high vowels adjacent to semivowels, or very reduced schwas. Further work should include improvements to the output confidence score, and error correction as part of vowel quality detection.

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

Acknowledgements

My first and foremost acknowledgements go to Professor Ken Stevens, who has taught me more than anyone else about speech science, and devoted his time and effort to advising this thesis. He has been an inspiration and role model for me, an example of insight, patience and understanding to me and to generations of students, and he has fostered the laboratory and community of scientists where I have learned and grown. He also provided the funding which made this work possible. Thank you, Ken – for all these reasons, I couldn't have done it without you.

Thanks to my thesis readers, Dr. Jim Glass and Professor Carol Espy-Wilson, who have been invaluable sources of insight and critical advice, and shaped both my work and my outlook. Special thanks and remembrances to my third reader, Professor Jon Allen, who offered me valuable instruction and counsel before his passing in April 2000. Jon will be missed by all of us in the Research Laboratory of Electronics.

Many of the staff and students of the Speech Communication Group have helped me with information and suggestions large and small. Stephanie Shattuck-Hufnagel, Sharon Manuel, Elizabeth Choi, Marilyn Chen, and all the members of the Lexical Access team have provided feedback and discussion. Arlene Wint and Seth Hall have kept the wheels turning. Majid Zandipour, Melanie Matthies, Lorin Wilde, and lots of others have offered encouragement and advice.

Many of my friends have given me vital assistance (and, when necessary, distraction!). My lover Lee Steele has been a wonderful source of support and encouragement. My friends David, Lyman, Marek, and many other of my friends and family have all helped me stay healthy and sane through this work.

Last and most important, I offer my deepest thanks and blessings to Mother Earth and Father Sky for the world They carry between them, and the strength and wisdom that underlies all we do. I also thank and bless my spirit guides, Dolphin and Horse, and all the powers great and small that weave through the world that is home to all of us.

This work was supported by a grant from the National Institute of Health (DC02978).

Contents

1	Introduction: LAFF and Vowel Landmarks	16
1.1	Motivation	16
1.1.1	Thesis scope	17
1.2	The LAFF paradigm	18
1.2.1	Landmarks and Segments	19
1.2.2	Distinctive Features	20
1.2.3	Matching landmarks to the lexicon	21
1.3	Issues facing Robust Speech Recognition	21
1.3.1	Sources of speech variability	21
1.3.2	Lexical vs. acoustic representations	23
1.3.3	Statistical vs. knowledge based recognition	26
1.3.4	Modular vs. integral architecture	28
1.4	Theoretical and practical implementation issues	29
1.4.1	Definition of a Vowel landmark	29
1.4.2	Design goals	30
1.4.3	Optimization criteria	31
2	Background of Automatic Syllable Detection (ASD)	35
2.1	How Vowel landmarks relate to ASD	35
2.2	Early work, mostly knowledge based	37
2.3	The ARPA SUR project and its aftermath	38

2.3.1	Results of the ARPA SUR project	45
2.3.2	Impact of the ARPA SUR project	46
2.4	Recent work, mostly statistically based	48
2.5	Summary	50
2.5.1	Detectable characteristics of vowels	50
3	Statistical Study of Database Vowels	53
3.1	Predictions of F1 behavior in vowels	53
3.1.1	(1) Presence of F1 peaks in the vowel	55
3.1.2	(2) F1 amplitude and frequency peak together	55
3.1.3	(3) F1 peak is better than midpoint for vowel quality	55
3.1.4	(4) F1 peak can be approximated without formant tracking	56
3.2	Corpus	56
3.3	Methodology	56
3.4	Signal processing	57
3.4.1	Formant tracking	57
3.4.2	Amplitude computation	59
3.4.3	Peak detection	60
3.4.4	Basic data	61
3.5	Experiment 1: Presence of F1 Peak in Vowels	68
3.5.1	Methodology	69
3.5.2	Experiment 1A : Amplitude peaks against context	70
3.5.3	Experiment 1B: Frequency peaks against context	77
3.5.4	Conclusions to Experiment 1	91
3.6	Experiment 2: Coincidence of Amplitude and Frequency Peaks in Vowels	91
3.6.1	Methodology	92
3.7	Experiment 3: Vowel quality better at F1 peak than midpoint	104
3.7.1	Methodology	105
3.7.2	K Nearest Neighbors (KNN) Classification	107

3.8	Experiment 4: Fixed energy band is comparable to formant tracking	113
3.8.1	Methodology	114
3.8.2	Conclusions	123
3.9	Conclusions of the statistical study	123
3.9.1	(1) Presence of F1 peaks in the vowel	123
3.9.2	(2) F1 amplitude and frequency peak together	124
3.9.3	(3) F1 peak compared to midpoint for vowel quality	125
3.9.4	(4) F1 peak can be approximated without formant tracking	125
4	Vowel Landmark Detector (VLD) Implementation	126
4.1	The Baseline VLD	127
4.1.1	Front end processing	128
4.1.2	Feature extraction	128
4.1.3	Detection	129
4.1.4	Post processing	131
4.2	Experimental Issues	131
4.2.1	Failure modes	131
4.2.2	Scoring technique	132
4.2.3	Corpus	134
4.3	Baseline Experiment	135
4.4	Combination of Acoustic Measurements	138
4.4.1	Using intuitive nonlinear combinations	139
4.4.2	Using Neural Networks	139
4.5	Incorporating Neural Network into Vowel Landmark Detector	147
4.5.1	Issues	147
4.5.2	Implementation and validation	150
4.5.3	Training algorithm	150
4.5.4	Training results	151
4.6	Error characterization	155

4.6.1	Canonical error categories	157
4.6.2	Modified error categories	158
4.6.3	Additional modification: Skewed detections in consonants	164
4.6.4	Conclusions	169
4.7	Examples from the TIMIT database	169
4.8	Conclusions	174
5	Implications and Future Work	175
5.1	Enhancements to Vowel Landmark detection	175
5.1.1	Further improvements for VLD	175
5.1.2	Error characterization	176
5.1.3	Adaptability to other databases	177
5.1.4	Vowel classification schemes	178
5.2	Confidence scores	179
5.2.1	Generation of confidence scores	179
5.2.2	Validation of confidence scores	180
5.2.3	Example: Hybrid confidence score	181
5.3	System integration issues	182
5.3.1	Optimization criteria	182
5.3.2	Lexical contact and error recovery	183
5.4	System design issues	185
5.4.1	Acoustic cues and Information content	186
5.4.2	Representation of Distinctive Features	187
5.4.3	Calibration of Feature Values	188
5.4.4	Landmark and Feature Hierarchy	190
5.4.5	Lexical Matching and Phonetic Rules	191

List of Figures

1-1	Receiver Operating Characteristic. The curve shows the performance of a Likelihood Ratio Test as a function of its threshold value. Different threshold values can be chosen by different decision rules (dotted lines).	32
3-1	Vowel duration histogram	62
3-2	Token Counts by Sentence Position	65
3-3	Amplitude Statistics by Sentence Position. A slight downward trend is noticeable for all positions except the first (at zero).	66
3-4	Vowel Amplitude Histogram by Sentence Position	67
3-5	Beginning Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak at the beginning, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.	82
3-6	Middle Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak in the middle, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.	83
3-7	End Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak at the end, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.	84
3-8	F1 Amplitude Peak Histogram. The data include all vowels in table 3.18. The horizontal axis is the amplitude peak location, normalized against the duration of the vowel. The histogram is dithered to avoid crenellation artifacts.	94

3-9	F1 Frequency Peak Histogram. The data include all vowels in table 3.18. The horizontal axis is the frequency peak location, normalized against the duration of the vowel. The histogram is dithered to avoid crenellation artifacts.	95
3-10	F1 Amplitude and Frequency Peak Cross Histogram. The data include all vowels in table 3.18. The horizontal axes are amplitude and frequency peak locations, respectively, normalized against the duration of the vowel.	96
3-11	Schematic of expected histogram of F1 peaks on the difference diagonal.	97
3-12	F1 Peak Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin.	98
3-13	F1 Peak Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, in milliseconds. The vertical axis is the number of tokens per bin.	99
3-14	Schematic of expected assumption violations among F1 peaks on the difference diagonal.	100
3-15	Assumption Violation Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the average number of assumption violations per token, for tokens in that bin.	101
3-16	F1 Peak Sum-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the sum of frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin.	102

3-17	F1 Peak Sum-Diagonal Histogram, Voiceless and Voiced Context. The data include all vowels in table 3.18 which are preceded by stop consonants. The horizontal axis is the sum of frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin, computed separately for vowel tokens preceded by voiceless and voiced stop consonants.	103
3-18	Trapezoidal window for weighting.	115
3-19	Performance as a function of LowerFrequency	118
3-20	Performance as a function of UpperFrequency	119
3-21	Performance as a function of LowerWidth	120
3-22	Performance as a function of UpperWidth	121
4-1	Convex hull algorithm (after Mermelstein). See the text for a description of the procedure.	129
4-2	Experiment 5 MLP Performance by number of hidden units	142
4-3	Experiment 5 MLP network weights. These weights result from training using back propagation on a sum-of-squares error criterion. See the text for interpretation of the values.	144
4-4	Hyperbolic tangent, which is used as saturating nonlinearity for MLP units.	145
4-5	MLP Training – Distance from Starting Point	153
4-6	MLP network weights, final version. These weights result from training using gradient descent on an error rate criterion. See the text for interpretation of the values.	156
4-7	Histograms of percent vowel deletions as a function of sentence position. The horizontal axis is the vowel count, counting from the beginning of the sentence(upper) and from the end of the sentence (lower).	165

4-8	Scatter plot of skewed detections in consonants, plotting the time between the landmark and the segment boundary against the duration of the consonant segment. Most of the segments are fairly short, and most of the landmarks are close to the segment boundary.	168
4-9	TIMIT Vowel-semivowel example. The sentence is SX9 "Where were you while we were away" uttered by male talker PCS0. Skewed detections occur at 0.64 s ("you") and at 1.04 s ("we"), and two-sided skewing occurs at 1.28 s and 1.42 s (first two vowels in "were away").	171
4-10	TIMIT Vowel-vowel example (page 1). The sentence is SX172 "The triumphant warrior exhibited naive heroism" uttered by female talker EAR0. VV deletions occur at 0.60 s (the second vowel in "triumph-") and 1.40 s (the second vowel in "warrior" and the first vowel of "exhibited"), each appears as a shoulder on the adjacent vowel.	172
4-11	TIMIT Vowel-vowel example (page 2). The sentence is SX172 "The triumphant warrior exhibited naive heroism" uttered by female talker EAR0. Correct detection of two vowels in sequence occurs at 2.30 s ("naive") and epenthetic insertion of a vowel occurs at 3.00 s (final nasal in "heroism"). . .	173
5-1	Translation scheme for Distinctive Features, showing how continuous values (derived from acoustics) relate to discrete values (represented in the lexicon). The acoustic Value is plotted on the vertical axis, and the acoustic Confidence is plotted on the horizontal axis.	189

List of Tables

2.1	Knowledge Based Syllable Detectors	39
2.2	Goals and Final (1976) System Results for the ARPA SUR Project	46
2.3	Statistically Based Syllable Detectors	49
3.1	Experiment vowel labels and categories.	58
3.2	Experiment vowel categories, counts and mean durations.	61
3.3	Crystal & House data on vowels, counts and mean durations. This figure reprints the vowel duration data from the publication, Table 1. Syllable duration data are not reprinted here.	64
3.4	Experiment 1 classes for vowel contexts.	69
3.5	Experiment 1A statistical results. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.	71
3.6	Experiment 1A assumption violations, labeled by hand.	73
3.7	Experiment 1A assumption violations, labeled automatically.	74
3.8	Experiment 1A residual assumption violations, labeled by hand. The Talker and Sentence identify the utterance, and the Index identifies the vowel in question, numbered sequentially from the beginning. The vowel length is in milliseconds. The vowel in question is capitalized in the Orthographic fragment.	76

3.9	Experiment 1B statistical results. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.	78
3.10	Experiment 1B statistical results, female talkers only	80
3.11	Experiment 1B statistical results, male talkers only	81
3.12	Vowel height classes. Schwas, diphthongs, and syllabic sonorants are not included, because their acoustic manifestation of height is liable to be uncertain or ambiguous.	81
3.13	Experiment 1B statistical results, by vowel height	82
3.14	Experiment 1B statistical results for truncated vowels. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.	87
3.15	Experiment 1B context statistics, for frequency peaks at beginning of segment. Columns show the class of the vowel token, and rows show the manner of the preceding segment. Each entry shows the token count followed by the percent in parentheses.	88
3.16	Experiment 1B context statistics, for frequency peaks at end of segment. Columns show the class of the vowel token, and rows show the manner of the following segment. Each entry shows the token count followed by the percent in parentheses.	89
3.17	Experiment 1B assumption violations, labeled automatically.	90
3.18	Experiment 2 vowel categories and counts. For each category, these data are counts of all vowels with amplitude and frequency peaks both in the middle of the vowel.	92
3.19	Experiment 3 vowel categories and counts.	106

3.20	Experiment 3 vowel recognition rates by height, using non-interpolated formant tracks. Diphthongs are not included in these statistics.	110
3.21	Experiment 3 vowel recognition rates by height, using interpolated formant tracks.	111
3.22	McNemar Test results. The measurements made at the F1 peak are separated on the vertical axis, and the measurements made at the midpoint are separated on the horizontal axis.	112
3.23	Experiment 4 statistical results, for the canonical parameters. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.	117
4.1	Parameters for Vowel Landmark Detector, with typical values.	127
4.2	Scores by frequency range, with and without fricative detection. The “broadband” condition is Mermelstein’s original frequency range (500 Hz - 4 kHz), and the “F1” range is 0 - 650 Hz.	137
4.3	Scores by vowel stress. In general, less stressed vowels are more difficult to detect. The exception is lax vowels, which are easier to detect in context (because they are always followed by consonants).	137
4.4	Basic statistics for the Vowel Landmark Detector, using very lenient parameters to minimize deletion errors.	141
4.5	Training results for the VLD using MLP for decisions, for the first eight annealing runs. Token Error Rate (TER) is as defined in section 4.3. Distances are Euclidean distances in coordinate space.	152
4.6	Training results for the VLD using MLP for decisions, showing the ten points with best TER performance, out of 64 training runs.	154
4.7	Training results for the VLD using MLP for decisions, showing the results of nine annealing runs based on point A from Table 4.6.	155

4.8	Test results for the final VLD using canonical error categories. Percentages are relative to vowel count. Error rate is insertions plus deletions.	157
4.9	Test results for the final VLD using modified error categories. Percentages are relative to vowel count. Error rate is all insertions plus simple deletions.	159
4.10	Category statistics of detected vowels, using the same vowel categories as in table 3.1.	161
4.11	Manner characteristics of segments with LM insertions.	161
4.12	Statistics of stops with LM insertions. The percentages are relative to the total number of LM insertions in stops.	162
4.13	Statistics of fricatives with LM insertions. The percentages are relative to the total number of LM insertions in fricatives.	163
4.14	Vowel categories of deletion errors, using the same vowel categories as in table 3.1.	163
4.15	Test results for the final VLD using modified error categories, including skewed detection in consonants. Percentages are relative to vowel count. Error rate is all insertions plus simple deletions.	166
4.16	Manner characteristics of LMs for skewed detections in consonants.	167
4.17	Vowel categories of vowels which show skewed detections in consonants, using the same vowel categories as in table 3.1.	167

Chapter 1

Introduction: LAFF and Vowel Landmarks

This chapter is an introduction to the LAFF paradigm and the concept of Vowel landmarks. It includes discussion of the problems which a Vowel landmark detector must face, methodologies for testing and validation of a Vowel landmark detector, and the basic theory that underlies the definition of Vowel landmarks.

1.1 Motivation

A primary motivation for this thesis is to detect Vowel landmarks as part of the front end of a LAFF speech recognition system. LAFF [84] is a knowledge based approach to speech recognition, in which landmarks (indicating vowels, consonants, or glides) are detected in the speech signal, and phonetic features are detected and attached to the landmarks. Therefore,

landmark detection is a crucial first step in LAFF processing. Landmark detectors for Consonants [54] and Glides [86] have already been developed, leaving only Vowel landmarks yet to be done.

In addition, there are many other uses for automatic syllable detection, which is a task very similar to detection of Vowel landmarks. Among them are visual speech aids for the hearing impaired [37], database labeling aids, tools for perceptual studies, and automatic detection of rate of articulation.

1.1.1 Thesis scope

The goal of this thesis is to create a Vowel landmark detection algorithm that is simple and reliable, and to test and validate its performance on a standard database of continuous speech (the TIMIT database [47]). The algorithm should be able to accept speech by adult talkers of either gender, in all the dialects of American English represented in the database.

One of the long term goals of the LAFF paradigm is a speech recognition system which is insensitive to the production characteristics of the input speech. For this purpose, the system should be tested on multiple databases, including spontaneous as well as read speech, under differing conditions. Other researchers, notably Liu [54], have focused on this aspect of development. In the interests of time and complexity, this thesis will use only one database, leaving input invariance as a task for later consideration.

1.2 The LAFF paradigm

LAFF (Lexical Access From Features) is a proposal for a model of speech recognition by humans, intended to reflect how a human listener takes in the speech signal and derives a sequence of words from it. As such, it is not primarily a proposal for a commercially viable speech recognition system (although there is some argument that since speech is created by humans for humans, modeling the human process of perception is the most reasonable approach for an automatic speech recognition system).

Linguistic science has established that words are stored in human memory as sequences of segmental units called phonemes (along with a small amount of additional information, indicating stress level and other prosodic information). It is also known that each phoneme is stored as a collection of Distinctive Features (DFs) which are canonically represented as binary variables.¹ Furthermore, there is evidence that the DFs are not a disordered aggregate, but are grouped in an hierarchical structure that reflects the physical properties of the speech articulators [43].

In contrast to the segmental nature of phonemes and DFs, the acoustic speech signal is continuous in nature. Attempts to separate the speech signal into phoneme segments, and then to detect DFs within each segment, have met with only very limited success, due to the variety of phenomena which transform the discrete phonemes and DFs into the continuous signal.

¹Almost all DFs are allowed to be unspecified in value, as well as taking on plus and minus values, so that strictly speaking they are trinary variables. This is important in implementation, but does not affect the structure being presented here.

1.2.1 Landmarks and Segments

The primary innovation of LAFF is the notion of landmarks. The problem of segmentation of the speech signal is avoided by using landmarks rather than segments to break the signal into an ordered sequence of objects that can carry DF information.

The most fundamental distinction among phonemes is between vowels and consonants. Vowels are produced with the vocal tract fully open, causing the characteristic pattern of formants, while consonants involve closing the vocal tract to some degree. Consonants may be further divided into abrupt and nonabrupt consonants. Abrupt consonants are produced with a constriction that is strong enough to cause acoustic discontinuity (at least in some region of the spectrum), while nonabrupt consonants are produced without such a severe constriction, and hence do not exhibit acoustic discontinuity. Nonabrupt consonants are represented with Glide landmarks, while Consonant landmarks implicitly represent abrupt consonants.

So there are three classes of landmarks: Vowel, Glide, and Consonant. Most landmarks (though not all, see below) are located at a specific event in the acoustic signal. A Vowel landmark, for instance, is located at the maximum of low frequency energy in the vowel. An intervowel Glide landmark is located at the minimum of low frequency energy between two vowels. In these cases, each landmark corresponds to a single underlying segment.

Consonant landmarks are located at the closure and release of the (abrupt) consonant. In general, therefore, there are two Consonant landmarks which correspond to the underlying consonantal segment. However, it is possible for only one Consonant landmark to appear in correspondence to the underlying segment, typically in consonant clusters.

There are also cases in which a landmark is not located at a specific acoustic event. A pre-

vocalic Glide which appears after a Consonant and before a Vowel, for example, corresponds to an underlying glide segment, but is not located at an acoustic event. Similarly, some Vowel landmarks will be located where there is no peak in low frequency energy (typically in vowel-vowel sequences). Such a landmark, which is not located at an acoustic event, is a “floaters” which appears somewhere between the events of the landmarks which precede and follow it. The landmarks are always understood to have a fixed order in time.

This means that not all vowels will have a landmark that is generated from an acoustic event. In this thesis, the assumption is that all vowels have landmarks, but not all can be derived from acoustic events alone. Some landmarks may be generated by subtler acoustic information, such as formant movements. This assumption, and the conceptual distinction between “acoustic event” and “landmark” may not be shared by other researchers working on the LAFF project [54], [12].

1.2.2 Distinctive Features

For each class of landmark (Vowel, Glide, and Consonant), there are distinctive features² (DFs) which should be assigned to them. DFs are determined by measuring acoustic properties in the vicinity of the landmark and combining the acoustic properties according to rules. The details of the DFs are rules which are described elsewhere [84] and will not be covered here.

²The term “distinctive” refers to linguistic features which can make a distinction between phonetic segments. The phrase “distinctive feature” is used here to distinguish these features from acoustic features used in recognition.

1.2.3 Matching landmarks to the lexicon

Once the landmarks have been detected and populated with features, they must be matched to the words in the lexicon. Since words are stored in the lexicon as sequences of phonemes, a sequence of phoneme segments (perhaps more than one) must be posulated from the sequence of landmarks. The details of this process are not yet clear. The process will include phonetic rules for transformation of features and similar phenomena. See the literature [84] for more information.

1.3 Issues facing Robust Speech Recognition

There are several issues that complicate the task of robust speech recognition. In general, these issues pertain to speech recognition systems as a whole. Since a Vowel landmark detector is an important part of a speech recognition system, a review of these issues is relevant to this thesis.

1.3.1 Sources of speech variability

There are many sources of variability which can challenge a speech recognition system. We may group them into the following categories, in roughly increasing order of the difficulty they pose for robust recognition.

Additive noise includes background noise from the channel (such as white or pink noise) as well as background noise from the environment. Environmental background noise is either nonspeechlike (frequently encountered in an aircraft cockpit or automobile) or speechlike

(background babble, as in an office or other crowded environment). Additive noise may be stationary (background hiss, or steady vehicle noise) or nonstationary (pops, crackles, honking horns, etc.) but it is always uncorrelated with the speech signal.

Convolutional noise includes reverberance or echoes, filtering, clipping or saturation of elements in the signal chain, and similar phenomena. Unlike additive noise, it is correlated with the speech signal, and consequently is more difficult to handle.

Talker variability can include both physical and habitual differences. Physical variability, such as the age, size, or gender of the talker, can affect the vocal tract length, F0 range, and the appearance of formants. Habitual variability, such as dialect and speech disorders (such as lisping) are stable for each talker, but can vary widely between talkers.

Production variability includes some of the most difficult phenomena to predict or to recognize. Talker stress and speech rate can have significant effects on speech production. Phenomena such as lenition, epenthesis, and coarticulation (see section 1.3.2) can drastically modify the acoustic characteristics which correspond to the underlying phoneme sequence. While linguistic knowledge can characterize many of these phenomena, the conditions under which they appear, and the degree to which they manifest, are often unpredictable.

Production variability is notorious for causing difficulty for speech recognition systems. In particular, the transition from isolated words to continuous speech is a great challenge, because variability (which often comes through overlapping of gestures) takes place across word boundaries as well as within words. Because such phenomena are especially prevalent in casual, continuous speech, robustness on continuous speech will be a central test of the vowel landmark detector in this thesis. Phenomena of production variability follow patterns which can be expressed via linguistic knowledge. Therefore, we hope that a LAFF system that can capture such linguistic knowledge will prove to be relatively robust on casual, continuous

speech.

1.3.2 Lexical vs. acoustic representations

A speech recognition system is typically evaluated by comparing its output to some sort of transcription of the input speech signal (which will be called *scoring* in this thesis, see section 4.2.2). If the transcription of an utterance were a straightforward, unambiguous representation of the information in the speech signal, transcribed databases would be easy to use for testing, but this is not the case.

Lexical transcriptions

Almost all speech databases include a *lexical* transcription, usually orthographic in nature, which represents the words being spoken. The syllabic nuclei of a lexical transcription (their number and locations) are generally simple and unambiguous. (We are not primarily concerned with syllable boundaries, which can be much more ambiguous.) Lexical transcriptions are also relatively easy to generate, which helps to account for their popularity.

Unfortunately, there are many cases where a lexical transcription is not a reliable indicator of the acoustic representation of syllable structure. For complete speech recognition systems (whose output is lexical in nature), this may not be a major problem. But feature extraction systems (like a Vowel landmark detector) attempt to represent the information in the acoustic signal, without reference to a lexicon, and differences between the acoustic information and the transcription are a very difficult problem to deal with.

Syllabic nuclei can be inserted by *epenthesis*, frequently when semivowels are next to stop

consonants (as in “please” becoming similar to “police”) or nasal consonants (“arm” becoming like “carom”). Epenthetic insertion of vowel-like sounds can also occur at the release of a final stop consonant.

Vowels can be reduced when unstressed, almost to the point of total disappearance (as in “support” becoming similar to “sport” except for aspiration) especially for reduced vowels in unvoiced contexts, such as the first vowel in “potato.” This type of reduction is called *vowel deletion, omission, or elision*.

Vowels in sequence with semivowels can be particularly hard to characterize. A word like “fear” should have one syllable, but it can be pronounced with two, and in extreme cases, the two vowels may appear to be separated by a glide, which is another kind of epenthesis. The situation is even more complicated with “fire” or “file” which are liable to be pronounced with one syllable in some regional dialects (American Southern) and two syllables in other dialects (American Northeast).

Two vowels in sequence can also be hard to characterize. A word like “coerce” (two syllables) may sound like “coarse” or “quirts” (one syllable), which is an example of *coalescence*, or it may have two well produced vowels together, or it may sound like “covers” with two vowels separated by an epenthetic glide. Coalescence also occurs across word boundaries, as in “see in.”

Some lexical transcriptions are time aligned with the acoustic signal, but most are not. Scoring is much easier with a time aligned transcription. Without it, the only practical alternative is to match the total number of syllables detected to the transcription, which does not allow identification of individual detections or errors. Worse, this scheme cannot identify a deletion in one area and an insertion in another area as an error (bearing in mind that a Vowel landmark indicates the existence of a syllabic nucleus, without characterizing

it).

For all these reasons, a lexical transcription is likely to be an inaccurate representation of the acoustic manifestation of syllable structure.

Acoustic transcriptions

Some speech databases include an *acoustic* transcription, which is usually a string of phones or allophones, representing the speech sounds manifested in the speech signal. Only the LAFF database [10] is transcribed directly as landmarks. Acoustic transcriptions are almost always time aligned, which helps in the scoring process.

One problem with acoustic transcriptions is that they are time consuming to generate. They require skilled transcribers who have to spend a fair amount of time creating the transcriptions, time aligning them, and cross checking to correct errors. Because of the effort involved, acoustic transcriptions are rather rare, especially for databases of spontaneous speech and casual conversations.

A more serious problem is that acoustic transcriptions are often not unique or unambiguous. A phonetic segment transcription imposes categorical decisions on acoustic information that varies across a continuum. How these decisions should be made is far from clear.

For instance, Pitrelli [68] describes the process for assigning phone boundaries in the aligned phonetic transcription of the TIMIT database.

For vowel-vowel transitions, which tend to be gradual, half of the total duration is allocated to each, unless one of the vowels is reduced, in which case the reduced

vowel is marked to be one-third of the total and the other vowel, two-thirds. For transitions between vowels and semivowels, one-third of the duration is allocated to the semivowel and two-thirds to the vowel. [ibid., p. 134]

While this procedure is unambiguous, it makes no attempt to examine the acoustic characteristics of the speech signal or to generate a transcription which reflects the acoustic evidence.

Even without such ambiguity, acoustic transcriptions are vulnerable to errors, because there is no easy way to check for consistency. Errors can be introduced by transcribers who have access to the orthography; lexical and semantic knowledge will also influence the transcription, even when the goal is supposed to be a purely acoustic representation. Different transcribers may also use different conventions in ambiguous situations, which would introduce even more inconsistency.

In sum, both lexical and acoustic transcriptions are likely to have problems when used as a reference to evaluate the performance of a syllable detector. The main problem with lexical transcription is the variability between the transcription and the acoustic information, while the main problem with acoustic transcription is availability. Section 4.2.2 will address methods for dealing with transcription problems.

1.3.3 Statistical vs. knowledge based recognition

Most of the early work in speech recognition used knowledge-based methods, attempting to explicitly incorporate speech knowledge. The expectation was that the insights of acoustics, speech production, and linguistic phenomena would aid in achieving good performance over a wide range of speech variations.

Statistically-based recognition methods have largely supplanted knowledge-based methods over the course of the last two decades. Although we have a fair amount of qualitative knowledge about speech, implementing this knowledge in quantitative algorithms is still very difficult. Statistically-based methods (aptly called “ignorance models”) avoid this problem, and have demonstrated better performance than knowledge-based methods for a variety of tasks.

Statistically-based methods do have their drawbacks. They generally impose a higher computational load on the system, they require time to train, and they do not generally allow the use of insight or speech knowledge in their design. None of these are major problems, except the following. The fundamental problem with statistically-based methods is their inability to cope with phenomena that are not well represented in their training data.

When the input data consist of clear, read speech, statistically-based methods can achieve very good performance, as shown by the results of the ARPA SUR project (section 2.3). Only recently have statistically-based systems been tested on casual, spontaneous speech (section 1.3.1), where they perform very poorly (see, for example, Lippmann [53]).

Casual speech, especially spontaneous speech, displays a wide range of variability (phenomena such as coalescence, epenthesis, and lenition) in a variety of combinations. Attempts to represent these phenomena (in all their possible combinations) lead to explosive growth of the training data set, which rapidly becomes impractical both to create and to use in training.

Knowledge-based methods, if they can be made to work well, should be more robust against the variability in spontaneous speech (because spontaneous speech is characterized by linguistic constraints that a knowledge based approach can incorporate, as discussed in section 1.3.1). This thesis will begin with a knowledge-based approach to vowel landmark

detection, with the understanding that some statistical methods may be added as appropriate, when necessary to achieve good performance (generally, when speech knowledge is not adequate).

1.3.4 Modular vs. integral architecture

Closely related, but separate, is the question of speech recognition (SR) system architecture. Statistically based systems generally have an integral architecture, with a single processing stage (such as an HMM) generating all the output, without explicit structure. Knowledge based systems generally have a modular architecture, with several stages of processing which pass information from one to another via clearly defined interfaces.

Modular architecture is a staple of algorithm design and software engineering in general. It allows each module to be developed, tested, and maintained separately, and even extended or replaced as better algorithms become available. When the interfaces between modules can be interpreted theoretically (e.g. as landmarks, features, and so on), a modular architecture is a natural for knowledge based processing.

The weakness of a modular architecture is its tendency to impose “hard” (irreversible) decisions on the early stages of processing. Any mistakes made by the first stage of processing will cause more severe errors in the processing stages that follow it. This problem is well known [71, p. 15] and will be called “cascade failure” in this thesis.

The primary way to avoid cascade failures, or at least to reduce their impact, is to minimize “hard” decisions in the early stages. In particular, measurements may be marked with a confidence score, which may be interpreted as a likelihood that the measurement is accurate. Maintaining confidence scores, and predicating later processing on the confidence of the

earlier measurements, helps avoid cascade failure. Therefore, the Vowel landmark detector should generate confidence scores for its output.

1.4 Theoretical and practical implementation issues

1.4.1 Definition of a Vowel landmark

The precise definition of a vowel landmark is set forth by Stevens [84] as:

The vowel landmarks are the places in the signal where there is a maximum in amplitude in the range of the first formant frequency. These places are usually where the first formant frequency is maximally high.

An accompanying figure [ibid., fig. 5.1] shows the first formant frequency tracks, as well as the relative amplitude of the first formant prominence. Choi [11, p. 5] states that “the vowel landmark is placed at a time when the amplitude of the first formant (F1) is a maximum.”

Thus, the definition gives several acoustic parameters (low frequency energy, first formant frequency, first formant amplitude) but does not explain how to combine them. As a practical matter, energy in a fixed low frequency band is a simple and reliable parameter, but accurate formant tracking is a notoriously difficult task (see, for example, [46]).

The definition also says “the time between the landmarks is rarely shorter than 120 ms, and, except when there is a pause, is rarely longer than about 350 ms in running speech.”

Although not a strict constraint, these times make it clear that duration information should be used, and give a rough outline of what duration values can be expected.

There is some uncertainty about the relationship between a vowel landmark and a syllabic nucleus. What is the difference? The quote above makes it clear that a vowel landmark is an *acoustically* defined entity. Some researchers seem to indicate that they consider a syllabic nucleus as a *lexically* defined entity, but this usage may not be consistent. This thesis will assume that syllabic nuclei are lexical in nature, and vowel landmarks are acoustic in nature (see the discussion in section 1.3.2).

This definition appears to assume that the presence of a vowel landmark is independent of the characteristics or quality of the vowel. That is, the vowel landmark detector (VLD) detects the vowel landmark without regard to whether the vowel is high, low, front, back, etc. Such an assumption is probably more valid for simple cases (a single vowel between obstruent consonants) than for more complicated cases (several abutting vowels in sequence). See section 2.5.1 for more discussion.

There is some question of whether the definition of a vowel landmark is language independent. For now, this thesis will be confined to American English (keeping in mind the possibility of dialectic variation, as discussed in section 1.3.2).

1.4.2 Design goals

Robustness against production variability

The most important goal of the VLD is robustness against production variability. The VLD should be able to deal with rapid, casual, continuous speech without major performance

degradation. (The definition of “performance” will be discussed in section 1.4.3.)

Robustness against cascade failure

The VLD must provide as much information as possible to following stages of processing. Each landmark consists of a time value (the location of the landmark in the speech signal) and a confidence score (the strength of acoustic evidence for a landmark, or confidence that this is a true landmark). The confidence score will help later stages of processing use the landmark appropriately.

Adaptability to input conditions

Changes in talker, talking rate, or level must not cause problems. The VLD should be insensitive to these changes or adapt to changing input conditions as necessary.

1.4.3 Optimization criteria

The VLD will be optimized or tuned for best performance. However, the criteria which define “best” will depend on the intended use of the VLD.

Detection theory describes several decision rules for a generic event detector. A Bayesian detector optimizes the expected cost of a decision, based on a priori probabilities of the event to be detected, and cost values for insertion and deletion errors. When a priori probabilities are not available, a minimax detector can be used, which minimizes the maximum average cost across all possible events. When cost values are not available, a Neyman-Pearson decision

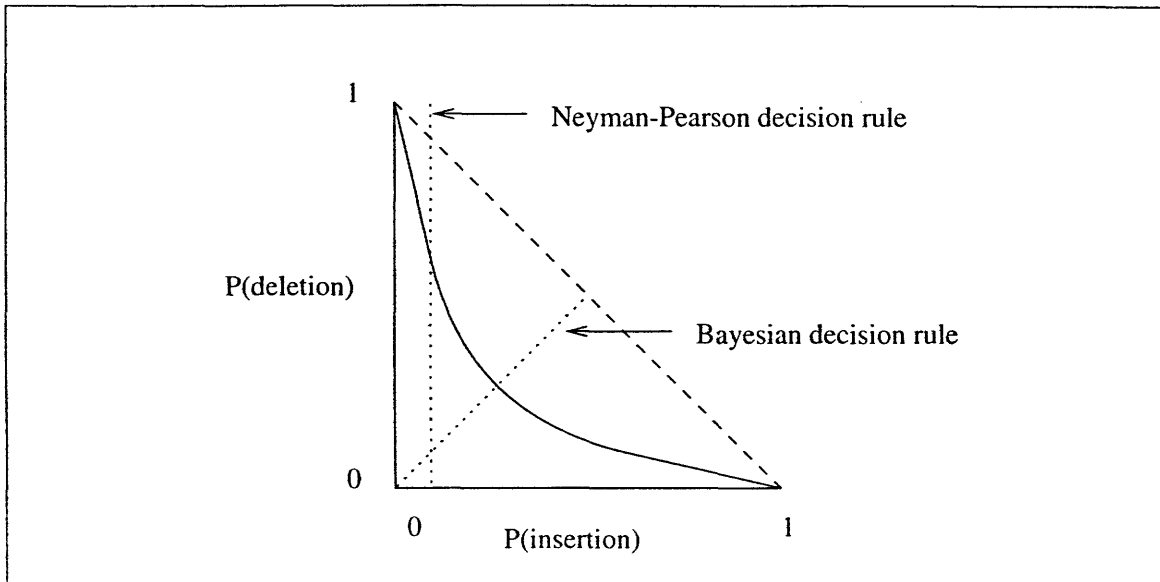


Figure 1-1: Receiver Operating Characteristic. The curve shows the performance of a Likelihood Ratio Test as a function of its threshold value. Different threshold values can be chosen by different decision rules (dotted lines).

rule can be used, which minimizes the probability of an insertion error while constraining the probability of a deletion error to a fixed value (or vice versa).

All these criteria can be implemented as special cases of a Likelihood Ratio Test, which subjects the measurement to a threshold (the threshold value depends on a priori probabilities and cost values). Its performance can be characterized by its Receiver Operating Characteristic (ROC) which shows the tradeoff between insertion and deletion errors as a function of the threshold value. See figure 1-1. The curve shows the performance of a Likelihood Ratio Test as a function of its threshold value. A Bayesian or minimax criterion corresponds to a line through the origin, as shown, whose slope is determined by the relative costs of insertion and deletion errors. A Neyman-Pearson criterion corresponds to a vertical line set by the desired probability of an insertion error (or, a horizontal line for deletion error).

A priori probabilities for Vowel landmarks, to the degree that they can be well defined (e.g. average frequency of landmarks in time), may be estimated from statistical studies of a training database of speech signals. Cost values, however, must be assigned based on the detector's application.

Criterion 1: Minimize error rate

The classic criterion for SR systems is minimum error rate, where the error rate is the sum of the insertion rate and deletion rate. Implicitly, this treats insertions and deletions as equal in cost. We will use this criterion for most (if not all) experiments, in order to achieve results which can be easily compared to other systems.

Criterion 2: Maximize information output

In order to avoid "hard" decisions, a different criterion should be used. "Soft" decisions require that as much information as possible is passed to the following stages of processing. This implies that the cost of a deletion error is much greater than the cost of an insertion error. Put another way, the VLD should output practically all its hypotheses for vowel landmarks, even those of very low confidence. As long as the confidence scores for the landmark hypotheses are available, this will allow the following stages of processing to decide which to keep (using other sources of information, such as phonotactic relations with consonant landmarks). This criterion is appropriate for applications where the VLD's output will be post processed by another module, for instance, integration with Consonant and Glide landmarks, using phonotactic knowledge.

Criterion 3: Maximize confidence in output

Cascade failure (section 1.3.4) is a concern for systems with modular architecture (characteristic of knowledge based systems). One way to avoid cascade failures is for the VLD to output only landmarks which have a very high confidence score. This implies that the cost of an insertion error is much greater than the cost of a deletion error. Put another way, the VLD should output only those landmarks which have very high confidence, so that following stages of processing do not get confused by false landmarks (insertion errors). This criterion is appropriate for applications where the VLD's output is not post processed or checked for consistency. In particular, this is a natural criterion for a multipass architecture, in which the first pass focuses on the most robust landmarks only.

For the purpose of the experiments in this thesis, we will use minimum error rate as the criterion, which is unambiguous and allows direct comparison with other systems. We will keep in mind that other criteria will probably be more useful when the VLD is integrated into a prototype system.

Chapter 2

Background of Automatic Syllable Detection (ASD)

This chapter is a review of past research and development of techniques for Automatic Syllable Detection (ASD) of human speech. Although there is very little published work describing Vowel landmarks as a concept, let alone algorithms for detecting them¹, ASD is a very similar task, with extensive research and development history.

2.1 How Vowel landmarks relate to ASD

Is Automatic Syllable Detection the same thing as Vowel Landmark Detection, and if not, how do they differ? This question depends on the circumstances in which a syllabic center (the presumed goal of ASD) is different from a vowel, or a vowel does not function as a

¹The only recent exception is Bitar [6].

syllabic center.

Recall the discussion in section 1.4.1, where it was assumed that syllabic nuclei are lexical in nature, and vowel landmarks are acoustic in nature. However, ASD as described in the literature always uses gross acoustic information to find syllables. The differences between lexical syllables and their acoustic manifestation (see section 1.3.2) are not discussed in published articles on ASD (see sections 2.2 and 2.4). This section will assume that the goal of ASD is to find the acoustic evidence for a syllable, which may not always correspond to the underlying lexical presence of a syllable.

Sonorant consonants /m, n, ng, l/ can appear as syllabic sonorant segments, usually manifestations of a schwa followed by a sonorant consonant. Syllabic sonorants can act as syllabic centers, as in the second syllable of “button,” but are not classified as true vowels. Presumably ASD will detect syllabic sonorants as syllabic centers. The published literature on LAFF [80] [84] do not make clear whether or not Vowel landmarks should be placed at syllabic sonorants. Their gross acoustic characteristics appear fairly vowel-like, with voicing and some formant structure, and they certainly fulfill the definition of a Vowel landmark, with a maximum in energy around the range of F1. A Vowel Landmark Detector would have to make some rather subtle measurements to differentiate syllabic sonorants from true vowels. For the purposes of this thesis, we will assume that syllabic sonorants should be marked with Vowel landmarks, just like true vowels, and if necessary, later stages of processing can be added to distinguish the landmarks of true vowels from those of syllabic sonorants.

In careful speech, it seems that all vowels function as syllabic centers. However, in casual, continuous speech, vowel-vowel sequences can be elided to various degrees, so that “naive” is pronounced like “knife,” or “coerce” becomes “coarse” for example. See section 1.3.2 for more examples of these phenomena, including examples of words whose syllable count seems to vary with dialect. Vowel-vowel and vowel-semivowel sequences deserve careful handling

for this reason. Section 4.2.2 will describe a scoring technique designed to deal with this kind of ambiguity.

In sum, it appears that ASD is similar enough to VLD that any differences are open questions rather than definite distinctions. Therefore, the published work in ASD is a reasonable starting point for VLD development. However, we want Vowel landmarks to be labeled with a meaningful confidence level or score, while most ASD techniques make a binary decision, without an output score. We will want to be aware of this requirement when choosing an algorithm.

2.2 Early work, mostly knowledge based

In the early days of speech recognition (SR), most of the techniques in use were explicitly knowledge-based, and syllable detection was usually one of the first steps in processing. Several leading research teams proposed schemes for automatic syllable recognition in the mid to late 1970's.

Weinstein and his team at MIT Lincoln Laboratories [88] and Kasuya and Wakita [41] both used LPC spectra to extract low-to-high frequency energy ratios. Weinstein detected syllabic nuclei by peak picking, while Kasuya used a conditioned linear discriminant function.

Mermelstein [61] and Medress et al. [58] both used power spectra to derive a low frequency energy profile, in the region of the first few formants of vowels. Medress detected syllabic nuclei by straightforward peak picking, while Mermelstein used a unique "convex hull" algorithm to recursively detect peaks which are prominent relative to their surroundings.

Zwicker et al. [93] used a critical band filter bank to derive low-to-high energy ratio profile, and detected syllabic nuclei by peak picking. Rabiner [70] used a normalizing technique on the total energy profile, and detected syllabic nuclei with a static threshold.

All of these techniques shared the basic premise of knowledge-based parameter extraction, followed by fairly straightforward peak detection. Some of them produced quite good results (see Table 2.1), but it is important to note that all were operating on high quality speech input: read speech in quiet, usually as isolated words or slow, carefully read sentences. The few experiments that used spontaneous speech (such as Pfitzinger et al. [69]) show substantially poorer performance.

If there is any insight to be gained from Table 2.1, it is that comparisons between different schemes are difficult to make. The wide variety of testing conditions and criteria make the performance numbers hard to compare directly.

2.3 The ARPA SUR project and its aftermath

By 1970, great progress had been made on talker-dependent recognition of isolated words from a limited vocabulary. Extending these techniques to multiple talkers, continuous speech, and large vocabularies was proving to be a very difficult task. Critics questioned the value of the speech recognition research being done [67].

In 1971, the Advanced Research Projects Agency (ARPA) initiated a massive research effort to develop speech recognition systems that would overcome these problems. Specifically, the goals of the Speech Understanding Research (SUR) project were stated as “Accept connected speech from many cooperative speakers in a quiet room using a good microphone with slight

	Front end	Parameters	Detection	Corpus	Error Rate
Mermelstein 1975	short term power spectrum	RMS energy 500-4000 Hz	convex hull recursion	slow careful speech in quiet	9.5%
Weinstein et al 1975	LPC 10 kHz SR 5 ms FP	1. F0 detection 2. RMS ratio LF/HF	dip detection peak picking	clear read sentences in quiet	3.0%
Medress et al 1978	FFT 10 kHz SR 10 ms FP	LF energy	peak picking and post proc for consonants		? (WER only)
Kasuya & Wakita 1979	LPC autocorr 10 kHz SR 12.8 ms FP	1. RMS energy 2. LF/HF ratio 3. LPC ratios	conditioned 2D linear discriminant		7.69%
Zwicker & Terhardt 1979	critical band filter bank	LF/HF ratio	peak picking (?)	German cities one talker in quiet	1.43%
Hunt et al 1980	mel-scale cepstral coefs	loudness profile	dip detection (with duration constraint)		? (SER only)
Rabiner 1984	RMS energy (zeroth order autocorrelation)	1. ST envelope 2. median smooth 3. normalize	static threshold		? (string error rate only)
Hermes 1990	Pitch synch FFT 5 kHz SR 10 ms FP	"Vowel strength" (sum of spectral peak factors)	peak picking with duration & level constraints	fluent Dutch Dutch words English words	11.1% 5.2% 4.7%
Fakotakis et al 1993		short term energy	peak picking with constraints		
Pfitzinger et al 1996	FIR bandpass 250-2500 Hz	log energy, smoothed, lowpass 9 Hz	peak picking with duration constraint	read German spon German	12.9% 21.0%
Bitar 1997	16 kHz SR 25.6 ms Hamming 5 ms FP	mid freq energy (two bands)	Mermelstein convex hull (sonorant only)	TIMIT read English in quiet	20.1%

Table 2.1: Knowledge Based Syllable Detectors

tuning/speaker accepting 1000 words using an artificial syntax in a constraining task yielding less than 10% semantic error in a few times real time on a 100 MIPS machine.” [45] The project lasted five years, involved hundreds of people, and resulted in the development of four major speech understanding systems as well as a number of supporting efforts.

All four systems were significant improvements over the previous generation of speech recognizers, but only Harpy, one of two systems designed at Carnegie-Mellon University, substantially met the SUR project’s goals. As will be described below, Harpy achieved its impressive performance by relying heavily on strong grammatical constraints, rather than improved front end algorithms such as syllable detection.

What follows is a more detailed examination of the four final systems, with particular attention to the front ends (initial stages of signal processing).

The SDC speech understanding system

The SDC speech understanding system [1] performed bottom-up analysis on the speech signal and produced an array of acoustic-phonetic data called the A-matrix, which was the only representation of the input used by the rest of the system.

The first task performed by the front end is to compute a pitch track [ibid, p. 273] . This is done in three stages: (1) the signal is lowpass filtered at 1000 Hz, (2) the autocorrelation spectrum is computed using a 50 ms window and peaks of the spectrum are picked as pitch candidates, (3) the pitch candidates are processed with median smoothing (to remove anomalous values) and RMS smoothing (to remove small irregularities). The second task of the front end is LPC analysis on 25.6 ms frames with a Hamming window, followed by peak picking and formant tracking. The next pass calculates the number of slope changes in the

digitized speech and marks dip areas, which will be used in segmentation. The next pass segments and labels sonorant regions that can be reliably handled. Besides the ARPABET phoneme labels, the system also used binary feature labels (such as retroflexion, nasalization, etc.) and rough labels for use in labeling ambiguous areas. Distance measures are computed from known points in a talker-dependent vowel table, areas of minimal change are located, rough segment boundaries are determined, and the distance measure is used to produce labels and scores. Another pass segments and labels non-sonorant regions (fricatives and stops) using special LPCs with short windows, matching the resulting spectra to stored templates. The next pass calculates prosodic information using intensity, duration, pitch, and other computations. Finally, each 10 ms frame of the speech signal is marked with a single "best" label. The resulting A-matrix contains all the information available to the rest of the system.

The rest of the system includes a top-down word hypothesizer with the ability to find coarse subsets, a detailed word verifier, and a phrase verifier with abutment rules. Lexical items are stored as directed graphs of phonemes; each path through the graph is an alternate pronunciation. The grammar is a set of context-free top-down production rules. The control system pursues a best-first search strategy.

The system as a whole recognized 30% of utterances exactly, with another 15% recognized as very similar (for instance, "is" substituted for "was") on the 1000 word vocabulary task, using a grammar with a branching factor of 105. The system was not trained to the talker's voice. This system was originally meant to work with a top end designed by SRI, which was not completed, forcing SDC to develop their own version, which was admittedly the weakest link in the project. Full performance evaluation of the final system was not done due to lack of funding. The experimenters, while recognizing that the top end (lexical access and grammatical engines) was the major weak part of their system, emphasize the importance of improving bottom end performance, noting "...there is more leverage to be gained at the bottom end for a given effort than almost anywhere else in the total system." [ibid, p. 290]

HWIM

BBN developed a system called "Hear What I Mean" (HWIM) [89] which operated by bottom up phonetic hypothesis by rule. The resulting hypotheses are arranged in a segment lattice with likelihood scores.

Speech was digitized at 20 kHz, preemphasized by first differencing and subjected to LPC analysis using 13 poles with a 20 ms Hamming window at 100 frames per second. Formant frequencies were estimated from the lowest bandwidth poles of the LPC results. Total energy, energy in low, mid, and high frequency bands, zero crossing rate, and spectral shape were also computed at 100 frames per second. All these parameters were directly available to higher levels for detailed top down word verification; in addition, the frame based parameters were used by a rule based system for bottom up phoneme hypotheses. The first rules to be applied performed very gross segmentation and classification (typically manner classification) and later rules operated on the results of the earlier ones by modifying segment boundaries, creating new segments, and modifying segment labels to generate a segment lattice. Finally, scores for the lattice segments were computed as likelihood ratios, based on confusion statistics and refined using acoustic analysis dependent on the segment label. [ibid, p. 322] The segment lattice allows multiple segmentation paths for cases where the acoustic evidence is not sufficient for unique segmentation decisions. However, there appears to be no explicit allowance for coarticulatory effects in the HWIM system, whereby one segment's label would depend on the labels for adjacent segments.

The bottom up hypotheses in the segment lattice are matched to a tree structured pronunciation dictionary. Some word boundary effects are captured in the dictionary by paths which link certain terminal nodes back to certain initial nodes in the network. For example, a word ending in /st/ can be pronounced without the final stop, but only when followed by a word beginning with /s/ [ibid, p. 323]. The word candidates generated by this bottom

up procedure are then verified against the frame-based acoustic parameters by top down synthesis by rule, using the Itakura distance metric and dynamic time warping. [ibid, p. 324]

An augmented transition network (ATN) captured the structure of the grammar. Syntactic and discourse information were incorporated after the initial lexical hypotheses were generated by an event-based control structure which built single-word matches progressively into larger parse structures using either left-to-right or island driven search strategies.

When tested in isolation, the front end of HWIM (everything up to the segment lattice) detected segment boundaries in the best path through the lattice with 1.7% deletion errors and 2.3% insertion errors. The correct label had the highest score about 57% of the time, but was among the top seven scores over 90% of the time. The entire system recognized 41% of utterances exactly, with another 2% semantically correct, on a vocabulary of 1097 words with a branching factor of 196. The experimenters note that HWIM's performance, unlike Harpy's, did not seem to be strongly dependent on high level constraints (vocabulary size and grammar branching ratio), and also that a great deal of tuning would have been possible with more time and funding, as opposed to Harpy which had been tuned to near optimum.

Hearsay-II

Carnegie-Mellon University developed a system called Hearsay-II [21] around a global data structure called a blackboard, which serves to integrate knowledge sources. Each knowledge source watched for a hypothesis (from some other knowledge source) to appear on the blackboard; and when it does, it applies its knowledge to write new hypotheses onto the blackboard (where they will be used in turn by other knowledge sources). Other knowledge

sources test the plausibility of extant hypotheses. In this way, diverse processes and sources of information interact through the data. This scheme offers great flexibility for developing and testing knowledge sources, alone and in various combinations.

Speech signals were digitized at 10 kHz, and a set of parameters referred to as ZAPDASH (Zero crossing rate And Peaks in Differenced And Smoothed waveforms [29]) was extracted. These parameters were used for segmentation by an iterative refinement technique: the signal was separated into silent and nonsilent regions, the nonsilent regions were separated into sonorant and nonsonorant, sonorant regions were separated into peak and nonpeak areas, and nonsonorant regions were separated into fricatives and flaps. The resulting segments (contiguous and not overlapping, i. e. not a lattice) were labeled by matching the central portion of the segment against phone templates using the Itakura metric, producing a vector of scores, one for each template. This rather crude labeling procedure is not justified in the literature; it may be that the experimenters did not feel that sufficient knowledge was available for the design of a more detailed scheme.

Bottom-up word hypotheses were generated by knowledge sources which parsed the labeled segments into syllables with a probabilistic production grammar, and grouped syllables into words according to a syllable-class dictionary. The words were scored with a top-down pronunciation network, adjusting their endpoints at the same time. Higher level knowledge sources embodied word sequence statistics (bigram model), phrase structure, semantic and discourse constraints, and halting criteria.

The experimenters claim 90% correctly or nearly correctly recognized utterances on a 1011 word vocabulary; however, this result was obtained with fairly extensive training and a restricted grammar whose branching factor was reported at about 33 [49, p. 69]. Apparently, high level task constraints can effectively overcome the handicap of simplistic front end processing for tasks of this magnitude.

Harpy

Harpy was also developed at CMU, and shared many features with Hearsay-II, particularly in the front end. Like Hearsay, Harpy used a sampling rate of 10 kHz, computed ZAPDASH parameters for segmentation, and detected segment boundaries using an iterative refinement procedure. The coefficients of LPC analysis at the midpoint of each segment were matched to speaker dependent templates using the Itakura metric. As in the Hearsay system, this front end is simplistic and fairly crude, discarding much spectral information. The experimenters acknowledge this, relying on language constraints to recover from errors [55, p. 353].

All the knowledge sources used in Harpy – alternate word pronunciation, word boundary effects, and finite state syntactic grammar – were compiled in to a massive network. This compilation is convenient because the rule systems are typically written for generation (top down). The resulting network of allophones is matched to the input segments via a beam search (best-few search). Although the network requires a large amount of memory, and any change in the lexicon or grammar would necessitate recompilation of the entire network, the computation is straightforward.

Harpy was the clear winner of the ARPA SUR “contest,” with 95% of utterances recognized as substantially correct on a 1011 word vocabulary with a grammar branching factor of 33. Even results with telephone speech were reported as favorable (increase in word error rate by a factor of 3 to 4). Harpy also was the most computationally efficient system.

2.3.1 Results of the ARPA SUR project

The results of the ARPA SUR project are summarized in table 2.2, after [49, p. 69]. Clearly, the best results are associated with constraining language models (low branching factor),

GOAL	RESULTS WITH 1976 ARPA SUR SYSTEMS			
	HARPY	HEARSAY II	HWIM	SDC
Accept continuous speech, from many cooperative speakers,	184 sentences 3 male, 2 female	22 sentences 1 male	124 sentences 3 male	54 sentences 1 male
in a quiet room, with a good microphone, with slight adjustments for each speaker,	20 training sentences	60 training sentences	no training	quiet room good mike no training
accepting 1000 words, using an artificial syntax, yielding less than 10% semantic error, in a few times real time (=300 MIPSS)	1011 BF=33 5% 28 MIPSS	1011 BF=33 or 46 9% or 26% 85 MIPSS	1097 BF=196 56% 500 MIPSS	1000 BF=105 76% 92 MIPSS

Table 2.2: Goals and Final (1976) System Results for the ARPA SUR Project

use of training sentences, and high computational efficiency. The CMU systems were also much more highly optimized and tuned than either HWIM or the SDC recognizer, which were displaying steady improvements up until the deadline with no sign of reaching peak performance.

2.3.2 Impact of the ARPA SUR project

“There exists a real danger that Harpy’s success may overshadow other equally important contributions of the ARPA SUR project [50, p. 390].” In reviewing the contributions of the project, most prominent scientists agreed that improvement of the initial signal processing performance was one of the most important directions for future research (despite the demonstrated utility of high level constraints for the ARPA SUR task). Lea ranks acoustic phonetic analysis first in priority for further work. “... The ‘front end’ analysis routines which transform from acoustics to phonetics, phonology, words, and prosodics are among the top priority components on which to focus.” [49, p. 86] In summarizing needs for future work, the same author notes “... the diversion that such higher-level modules have been, in detracting from needed work on the acoustic ‘front ends’ of recognizers [51, p. 563]” and “Primary attention should be on the front end ... of the recognition process, where the

critical gaps in current capabilities are known to exist [ibid, p. 567].”

Humans who listen to nonsense utterances (with no syntactic or semantic constraints) can recognize 85% of content words correctly, as opposed to 34% by Hearsay-II, which has the best performing word hypothesizer [78, p. 152]. This indicates that there is much more information in the acoustics alone than machine algorithms were using. Lea concludes his monograph by emphasizing the need for deeper understanding of acoustic phonetics once again:

Repeatedly throughout this book, I (and several other authors) have advocated work on the front-end of a recognizer. Yet, there is need for basic research projects to gather more necessary knowledge about characteristic features of continuous speech. This research will typically take years before it has direct impact on speech recognizers, so work should begin soon on these topics. [51, p. 568]

Nevertheless, Harpy’s dramatic performance for the ARPA SUR task spurred a great amount of interest in statistically based algorithms such as Hidden Markov Modeling and artificial neural networks. In a fairly short period of time, knowledge based systems (with or without automatic syllable detection) fell out of general use.

It is certainly true that knowledge based systems needed improvement. Several researchers examined Mermelstein’s convex hull algorithm around this time, and either rejected it as too error prone, or manually reviewed its output to correct errors [62], [56].

However, the outcome of the ARPA SUR project was far from a definitive endorsement of statistically-based methods over knowledge-based methods. First and foremost, its time limit allowed only a very short development cycle, which effectively biased its outcome in favor of a “quick fix.” Second, the speech input used was regular, clear read speech in

quiet, minimizing many of the phenomena (coarticulation, lenition and epenthesis) which characterize casual speech.

In any case, almost all of the SR work after the ARPA SUR project has been devoted to statistically based techniques. Since most such algorithms have an integral architecture (see section 1.3.4), separate syllable detectors were no longer required.

2.4 Recent work, mostly statistically based

Although not at the pace of the 1970's, work on syllable detection has continued in recent years. Much of the recent work has applied statistical recognition methods to the problem, following the general trend. Some of the recent work appears to be intended to add supplemental information to statistically based systems [91]. Other applications are visual aids for the hearing impaired [37], or studies of the talker's rate of articulation [15].

Reichl and Ruske [72] used log energy spectra (grouped into critical bands and normalized) as input to a multilayer perceptron, and produced good results in syllable detection on read German sentences, as shown in Table 2.3. Pfitzinger et al. [69], however, used low frequency energy, smoothed and filtered, and detected syllabic nuclei by peak picking, similar to older approaches.

The International Computer Science Institute has been directing effort towards incorporating syllable information (onsets rather than nuclei) into English SR systems. Shire [77] and Wu et al. [90] use auditory models (RASTA [36] and modulation spectrograms [33]) as input to a multilayer perceptron, in a particular effort to deal with spontaneous input (OGI Numbers95 database [14]).

	Features	Detection	Corpus	Token Error Rate
Reichl & Ruske 1993	FFT in critical bands 16 kHz SR 10 ms FP	ANN (MLP)	read German	5.2%
Green et al 1993	bark scale filter bank	HMM	TIMIT	? (SER only)
Hunt 1993	energy in CBs and cepstra	RNN	TIMIT	?
Shire 1997	1. RASTA PLP 2. Modulation spectrogram	ANN (MLP) three layer	spontaneous digits (Numbers95)	threshold 19.9% DP 11.7%
Wu et al 1997	1. RASTA PLP 2. Modulation spectrogram	ANN (MLP) three layer		21.0% (?)

Table 2.3: Statistically Based Syllable Detectors

Hermes [37], however, used a novel measure of spectral resonance peaks to produce a profile of “vowel strength.” Like Shire, he searches for syllable onsets rather than nuclei. The spectral peak measure requires pitch synchronous FFTs, but its performance is fairly good (see Table 2.1). Since his application is visual speech aids for the hearing impaired, low latency in real time is an important design goal. His scheme finds vowel onsets (not centers), justifying the more complicated signal processing.

Bitar [6] uses a syllabic feature² detector which is based somewhat on Mermelstein’s algorithm; however, it is used together with a sonorant feature detector. First the feature sonorant is estimated (its acoustic cues are periodicity and strong low frequency energy). Then the syllabic feature is sought only in regions which have been judged sonorant. Two energy bands are used (2.8 - 6.4 kHz and 2 - 3 kHz) and combined using a fuzzy logic algorithm. The recursive convex hull algorithm of Mermelstein is used to find peaks, but without secondary constraints (duration, absolute level, or fricative detection).

²Strictly speaking, Bitar’s algorithm is a detector for events associated with “syllabicity” which seeks the acoustic correlates of the phonetic feature [syllabic] [6, p. 132].

2.5 Summary

This survey shows that a fair amount of effort has been directed to the problem of automatic syllable detection over the last several decades. However, it is far from a solved problem. Few of the published authors compare their techniques to each other, and most of them use rather casual or *ad hoc* evaluation schemes, making it difficult to compare the performance of different systems.

2.5.1 Detectable characteristics of vowels

Review of the theory of acoustic phonetics, and of the published literature on automatic syllable detection, suggests that there are three different characteristics that indicate the presence of a vowel. The three characteristics proceed from the simplest in definition (but least accurate), to the most complex in definition (but most accurate).

The term “track” will be used here to mean the time course of a scalar parameter. For example, the first formant (F1) frequency is a scalar, measured at a particular point in time. The trajectory of F1 over time will be called the F1 track.

Here are the three detectable characteristics of vowels.

Prominence of a low frequency energy track

Prominence of a low frequency energy track is the characteristic used by Mermelstein [61] as well as most following work. This should separate vowels and sonorant regions between

obstruent regions, but not sequences of vowels and sonorant consonants. This is the simplest computationally and conceptually, but liable to perform poorly in sonorant regions.

Prominence of a formant-presence track

Prominence of a formant-presence track is the characteristic used by Hermes [37]. The author's intuition when reading spectrograms, that the presence of a vowel is most often characterized by clear formant structure, supports this idea.

This characteristic should separate vowels from most sonorant consonants (with the possible exception of some postvocalic liquids) and hopefully intervowel glides as well. Stevens [83] remarks that semivowels are characterized by at least two formants moving close together, so an ideal formant-presence measure should emphasize formants that are well separated. A detector for this characteristic will involve more computation than for the low frequency energy track, however. Hermes' system includes pitch synchronous spectral analysis, which requires a pitch detector, as well as several types of post processing.

Movements of formant tracks

Movements of formant tracks, especially F1, are theoretically appealing but difficult to implement. This includes Stevens' proposal of prominence of the F1 track, but might not be limited to prominences, and may include higher formants as well. This should detect practically all vowels, including abutting vowels in sequence. Because of the need for a formant tracker, however, this characteristic is the most involved computationally, and the most difficult to implement. Indeed, this characteristic has never been reduced to practice in any implementation of an automatic vowel or syllable detector.

This thesis proposes to implement and investigate the first characteristic (low frequency energy), and establish its performance capabilities and drawbacks. The second characteristic (formant presence) and the third characteristic (formant movements) are beyond the scope of this thesis, because of their computational complexity. It is possible to envision three stages of analysis which reflect these three characteristics, with each stage refining the output of the previous stage. To the degree that algorithms can be implemented for the second and third characteristics, they may be implemented and investigated in following work, to compare their performance to the first characteristic.

Observe that the discussion of the three characteristics has emphasized the importance of context for vowel detection (for instance, interobstruent vowels are easier to detect than intersonorant vowels, and abutting vowels in sequence are even more difficult). This thesis will also investigate the effects of context, in order to understand the issues involved with integration of the VLD into a speech recognition system. See section 5.1.4 for details.

Chapter 3

Statistical Study of Database Vowels

A formal definition of a vowel landmark is “a maximum in amplitude in the range of the first formant frequency. These places are usually where the first formant frequency is maximally high.” [84] As this definition is somewhat vague, we have performed a statistical study of the TIMIT database, searching for maxima in F1 frequency and amplitude. This chapter presents the predictions of acoustic theory, the experimental methodology, and analysis of the results of the experiments.

3.1 Predictions of F1 behavior in vowels

According to acoustic perturbation theory [83, p. 148], the first formant should be reduced in frequency whenever a constriction is made in the front half of the vocal tract. Since all English consonants are articulated in the front half of the vocal tract, we expect that all consonants will cause F1 to drop in frequency (in English). Conversely, F1 frequency is

maximally high when the front half of the vocal tract is maximally open, making the F1 maximum a natural choice for the location of the Vowel landmark.¹

Also according to basic acoustic theory, an increase in the first formant frequency should be accompanied by an increase in amplitude, both of the first formant itself and the overall spectrum [25, p. 55], assuming that the glottal source is fixed in amplitude and spectral shape. To the degree that this holds true, either F1's frequency or amplitude may be used to determine the Vowel landmark.

Two assumptions underlie these predictions. First, the glottal source should be fixed, without major changes in either amplitude or spectral content. Second, the influence of secondary articulation (in particular, nasalization and glottalization) should be minimal. A major change in the glottal source, or major influence of secondary articulators (such as nasalization), will invalidate the predictions.

There are also several pragmatic concerns when performing experiments. For instance, the database labeling may describe an underlying vowel, preceded and followed by underlying consonants, but may not correspond to the acoustic realization of the speech signal (for example, a nasal murmur instead of a vowel, or an elided consonant). Also, errors in the formant tracking algorithm will cause incorrect values for F1 frequency, which may also affect the computation of F1 amplitude.

The experimental studies on the database are intended to test the theoretical predictions. The experiments will be numbered as follows.

¹A possible exception occurs when semivowels like /r l/ are adjacent to high vowels like /i u/. In these cases, the source is liable to be reduced in the semivowel relative to the vowel, so that F1 amplitude may peak in the vowel even if F1 frequency does not.

3.1.1 (1) Presence of F1 peaks in the vowel

What percentage of vowels have a proper maximum in F1 (amplitude or frequency)? That is, how often does a maximum occur within the body of the vowel, and not at its endpoints? If a significant fraction of the vowels have only a degenerate maximum (at an endpoint), we will want to investigate aspects of how such vowels are produced, in search of violations of the assumptions which underlie the theoretical predictions.

3.1.2 (2) F1 amplitude and frequency peak together

How do the locations of the F1 amplitude maxima compare to those of the F1 frequency maxima? Basic acoustic theory predicts that they should occur together. If and when there are differences, we will look for reasons for the differences, as well as systematic effects dependent on context, or other factors.

3.1.3 (3) F1 peak is better than midpoint for vowel quality

Where do the F1 maxima appear (close to the center of duration, early, or late)? What is their distribution, and how does it depend on context? Although the definition of the Vowel landmark does not say anything about position, it appears that human labelers place the landmark around the center of duration.² We will look for patterns or systematic variations.

If the F1 peak occurs where the vowel is least influenced by adjacent consonants, we may

²This hypothesis arises from the author's observations of the LAFF database. Objective measurements are not easy to make, because the beginnings and ends of vowels are not labeled.

suppose that the F1 peak is a better point in time for vowel characterization than the midpoint of the vowel's duration (although the midpoint is fairly good for vowel characterization [38]). We will use an automatic vowel recognition scheme to perform a recognition task on both points and investigate the difference.

3.1.4 (4) F1 peak can be approximated without formant tracking

Can we find an energy band whose peaks more or less coincide with the peaks of F1 amplitude? This would mean a VLD that doesn't need a formant tracker. We will investigate how much performance is sacrificed by using a fixed energy band rather than a formant tracker, and attempt to characterize the places where differences are likely to occur.

3.2 Corpus

The TIMIT database was used for these experiments. The entire database was used (all utterances from all talkers and dialects, both training and test sets). Each utterance includes an aligned phonetic transcription file (*.phn) and an aligned word transcription file (*.wrđ) as well as the signal waveform itself (*.wav).

3.3 Methodology

Entropic ESPS tools were used to process and examine the files. In order to use these tools, the transcription files must be converted to the ESPS label file format. The ESPS tool

cnvlab(1-eps) does this. cnvlab(1-eps) generates error messages on some of the TIMIT transcription files, including (1) “first label doesn’t start at time 0” (2) “labels not contiguous” and (3) “WARNING: labels overlap.” None of the errors were judged to be severe enough to interfere with the experiment.

3.4 Signal processing

3.4.1 Formant tracking

The first step of signal processing was to extract the F1 frequency track for each utterance. The Entropic formant tracker called `formant(1-eps)` was used. This formant tracker generates formant frequency candidates by frame based LPC analysis of the waveform. The local costs of mapping LPC roots to formant frequencies are computed at each frame based on the frequencies and bandwidths of the component formants for each mapping. The cost of connecting each of these mappings with each of the mappings in the previous frame is then minimized using dynamic programming (a modified Viterbi algorithm). For a more detailed description of the formant tracking algorithm, see the Entropic documentation [20] or the published description [74].

Most of the parameters of the formant tracker were left at their default values, which were judged to be adequate for this experiment. However, the frame period was changed to 5 ms (from the default value of 10 ms) to allow finer time resolution of the peaks. First, the signal was downsampled to a sampling rate of 10 kHz, and highpass filtered at 80 Hz to remove low frequency rumble and hum. For each frame, the signal waveform was preemphasized (with a coefficient of 0.7) and then windowed with a 49 ms \cos^{*4} window (effective duration about 18 ms). On this data, a twelfth order LP analysis was computed by autocorrelation.

Labels	Category
ix ax ux axr ax-h	schwa
em en eng el	sonorant
ih eh ae ah uh	lax
iy ey aa ao ow uw er	tense
aw ay oy	diphthong

Table 3.1: Experiment vowel labels and categories.

Tracks for the first four formants were computed from the LPCs by minimum cost dynamic programming.

F1 was extracted from the resulting formant tracks, and its frequency track was filtered to isolate the vowels. First, all F1 frequency values which do not appear in labeled vowels were set to zero. The TIMIT labels which are considered vowels for this purpose are shown in table 3.1, column 1. Note that this table includes syllabic nasal consonants and syllabic laterals, because in most cases, the automatic syllable detector will be expected to find these tokens.

Second, F1 frequency values for all frames whose F1 bandwidth was greater than 300 Hz were set to zero. This was intended to eliminate obvious errors of the formant tracker. In particular, vowels which are adjacent to fricatives often have some fricative noise on the vowel side of the boundary. If this stage was not used, the F1 frequency track would often have erroneous peaks at these boundaries. The 300 Hz threshold was chosen after a manual inspection of several cases chosen at random from the database.

Of course, formant tracking is a notoriously difficult task, and this loose constraint on bandwidth could not eliminate all errors. Manual inspection of the formant tracks to screen out errors was judged impractical on a database of this size (about 80,000 vowel tokens). Instead, the experiments were done on the automatically generated formant tracks, and when unexpected results were found, manual inspection was used to ensure that the formant

tracks were not wildly inaccurate.

3.4.2 Amplitude computation

Once the formant frequency tracks were computed, they were used to direct the computation of the formant amplitude tracks. The formant amplitude was estimated by summing the energy of spectrogram bins in a narrow band around the formant frequency.

The first step was to compute the spectrogram using the ESPS tool `sgram(1-esps)`. For this task, the spectrogram used a 6.0 ms Hamming window with no preemphasis, and a 10th order FFT, at a frame period of 1 ms. The short window was intended to ensure a frequency resolution low enough to smooth out the individual harmonics.

Next, the spectrogram's spectral bins were smoothed in time. This operation was intended to ensure a time resolution low enough to smooth out individual pulses of the glottal excitation. Averaging was done across 15 frames (rectangular window) at 1 ms per frame. The average values were output every 5 input frames, resulting in output frames at 5 ms per frame (same as the formant tracker). The center frame of the 15 frame average was time aligned with the formant tracker's frame, so that the average values were synchronized with the formant tracker's output.

For each frame, the formant tracker's F1 frequency was used to compute F1 amplitude. The spectral bins were summed across a 100 Hz band centered on the F1 frequency. Since the summation was done in the magnitude squared domain, the resulting value is dominated by the strongest amplitude bin. In this way, the F1 amplitude is well approximated even if the F1 frequency has a moderate error. The resulting amplitude value was converted to dB. (The decibel computation was not calibrated to any particular reference value, because only

the relative values are needed to find the peak.)

3.4.3 Peak detection

Peaks of both frequency and amplitude values for F1 were found by a straightforward search. Peaks were found for each vowel in the aligned phonetic transcription (including all segments considered as vowels, as shown in table 3.1, column 1). The peak search was not constrained in any other way, and the resulting peak may be located anywhere from the beginning of the segment to the end.

In general, the time of the resulting peak was placed at the midpoint of the 5 ms frame of the formant tracker. Care was taken to handle the possibility of “flattops,” or peaks which span more than one frame with the same value. In this case, the time of the resulting peak was placed midway between the beginning of the “flattop” and its end.

There is a possibility of two different peaks of the exact same value in one vowel. It is extremely unlikely, given the floating point representation of frequency and amplitude, but possible. At the time, the algorithm would pick the first such peak, if there is more than one.

When these computations were complete, the resulting values were stored in a database for further processing. The database fields included the talker, the sentence, the number of the vowel in the sentence, segment labels for the previous and following segments (as well as for the vowel itself), start and end times for the vowel (in milliseconds), time and value of the frequency peak, and time and value of the amplitude peak. This database can then be conveniently analyzed to discover characteristics of the peaks, and is referred to as the “Raw” experimental data in subsequent sections.

Category	Count	Mean duration (ms)
Schwa	24314	62.3
Sonorant	2482	85.3
Lax	21398	100.8
Tense	27528	111.6
Diphthong	5134	151.8
All	80856	95.6

Table 3.2: Experiment vowel categories, counts and mean durations.

3.4.4 Basic data

The experiment on the entire TIMIT database yielded a total of 80856 vowel tokens. The first look at the data investigates how duration varies with vowel category.

Duration statistics

The mean duration for each category is shown in Table 3.2. Clearly schwa, lax, and tense are the most populous categories, while syllabic sonorants and diphthongs are far fewer in number. The categories are sorted by mean duration, schwa being shortest and diphthong being longest. Histograms of duration for each category are shown in Figure 3-1.

This figure clearly shows how much variation exists within each category, and how much overlap exists between categories. Duration information alone cannot separate any categories. In fact, duration information alone is not much help in determining the presence or absence of one or more vowels. From this data, we can only conclude the order of magnitude of vowel duration (from about 25 ms to 250 ms) that a vowel detector should be adjusted to look for.

The vowel duration data presented here compare well with published data on vowel duration,

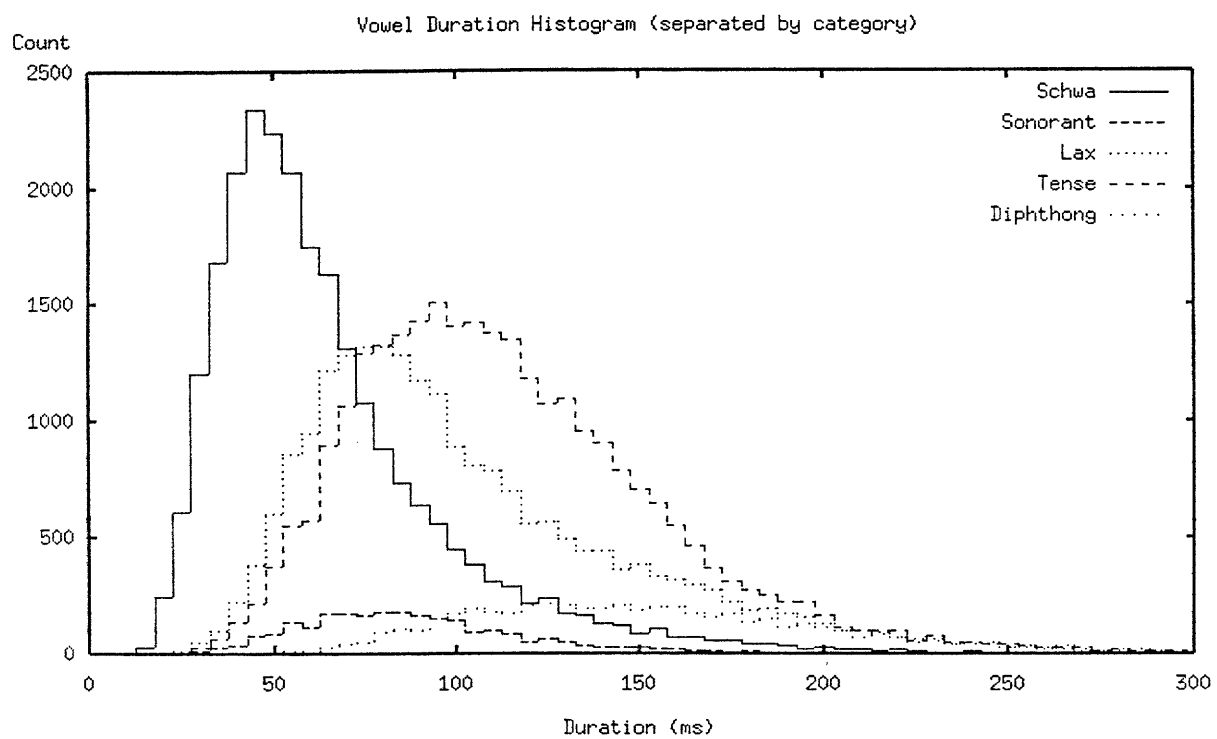


Figure 3-1: Vowel duration histogram

such as Crystal & House [15], a subset of whose data appears in table 3.3. (Vowel duration data are reprinted here, but not syllable duration data.) Crystal & House separate their data by stress, which is not available in the TIMIT labeling. However, the mean vowel duration for non prepausal syllables (127.6 ms stressed, 58.9 ms unstressed) is well within the bounds shown in Figure 3-1, while the mean vowel duration for the much less numerous prepausal syllables (185.9 ms stressed, 84.1 ms unstressed) is still reasonable. The mean duration for all vowels (103.8 ms) is very close to the mean duration for all vowels shown in Table 3.2 (95.6 ms, difference is about 8%).

Amplitude statistics

The amplitudes of vowels also carry information which is valuable to the process of landmark detection. In analyzing amplitude, it is important to note that amplitudes can vary substantially from one sentence to another, and that amplitudes tend to fall during the course of the sentence.

To examine amplitude phenomena, vowel tokens were grouped by their position in the sentence, and their amplitudes were normalized against the strongest token in the sentence. (This means that, after normalization, each sentence will have exactly one token whose normalized amplitude is 0 dB, and the other tokens will have values below it.)

The number of tokens in each sentence position is shown in figure 3-2. There are 80856 tokens in 6299 utterances, for an average of 12.8 vowel tokens per sentence. The figure shows that very few sentences have more than 20 vowels. None have over 24 tokens.

First and second order statistics for vowel amplitudes are shown in figure 3-3. The mean values decrease gradually over the course of the sentence (only about 0.35 dB per position

Syllable type	+ stress		- stress	
	N	ms	N	ms
	Not prepausal			
V	29	130	139	70
VC	93	127	383	64
VCC	12	133	8	60
CV	109	131	504	55
CVC	678	129	351	55
CVCC	354	126	11	52
CVCCC	13	113	0	-
CCV	34	133	0	-
CCVC	139	122	7	39
CCVCC	45	128	0	-
CCCVC	4	104	0	-
All	1510	127.6	1403	58.9
	Prepausal			
V	1	243	2	89
VC	19	197	20	95
VCC	6	135	0	-
CV	34	273	4	65
CVC	176	189	11	72
CVCC	47	149	1	84
CVCCC	3	104	0	-
CCV	1	308	0	-
CCVC	35	166	1	67
CCVCC	7	135	0	-
CCCVC	12	139	0	-
All	341	185.9	39	84.1

Table 3.3: Crystal & House data on vowels, counts and mean durations. This figure reprints the vowel duration data from the publication, Table 1. Syllable duration data are not reprinted here.

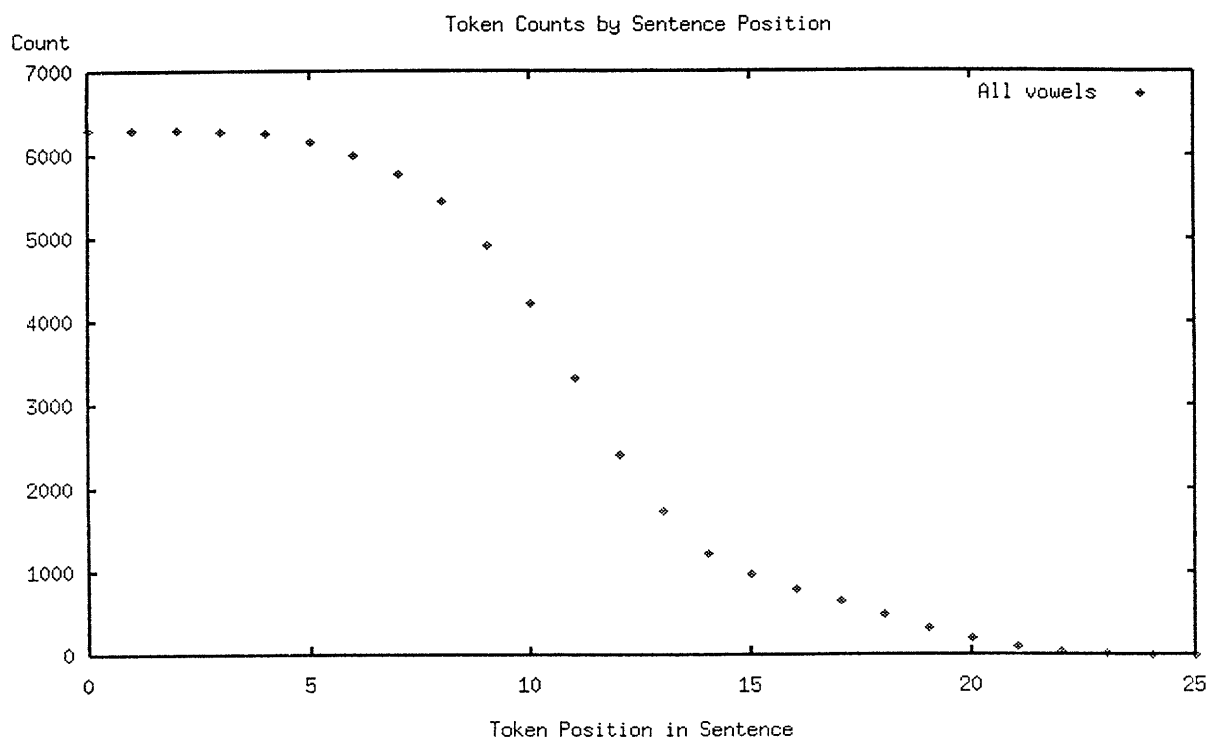


Figure 3-2: Token Counts by Sentence Position

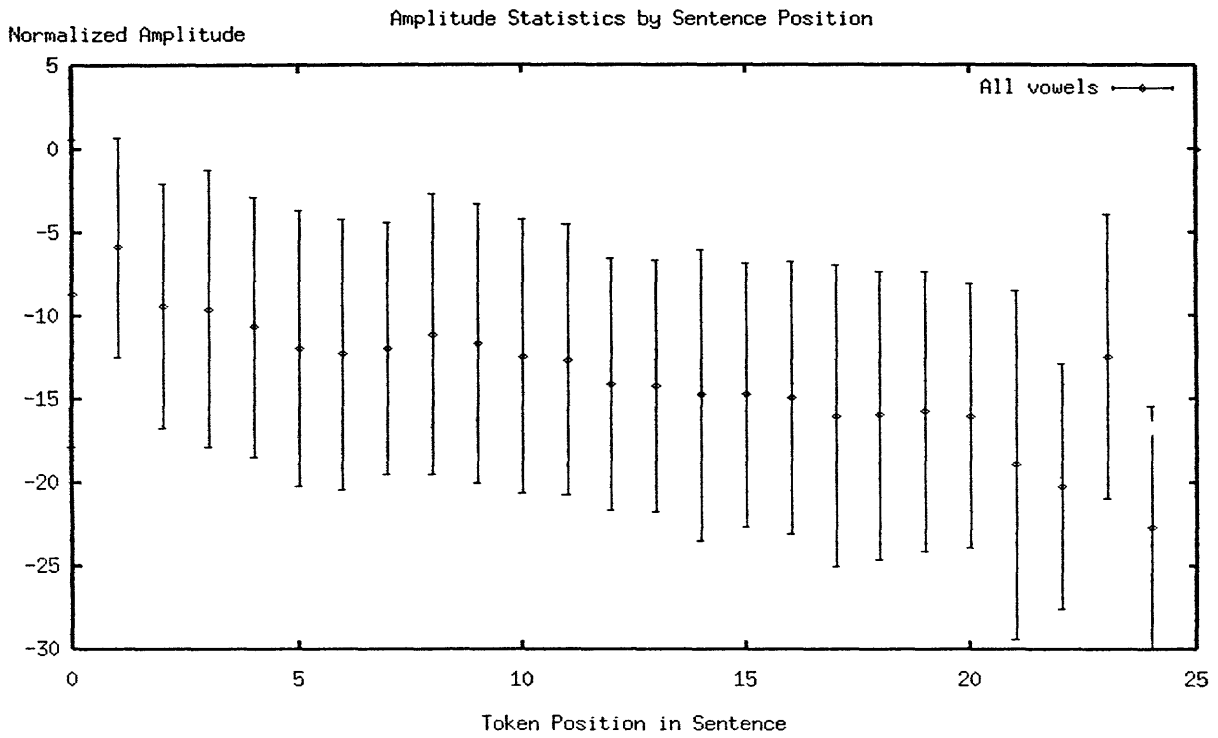


Figure 3-3: Amplitude Statistics by Sentence Position. A slight downward trend is noticeable for all positions except the first (at zero).

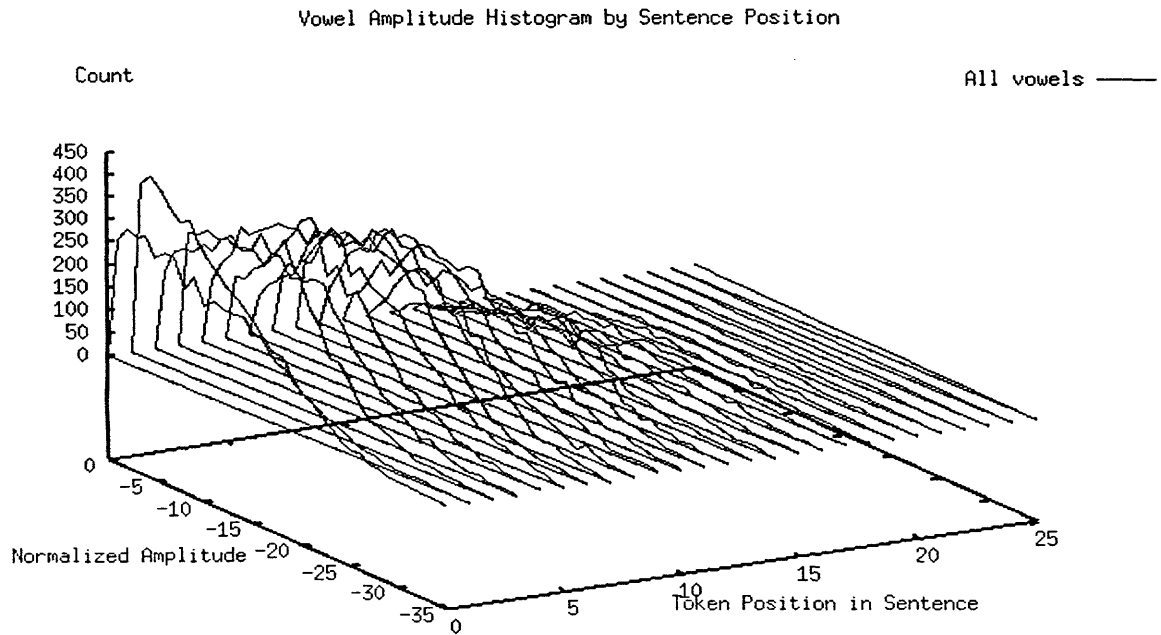


Figure 3-4: Vowel Amplitude Histogram by Sentence Position

increment) but the standard deviations seem to be relatively constant (7 to 8 dB). The statistics do show exceptions for the latest positions, but these are where the data are sparse (as shown in figure 3-2). Also, the very first vowel tends to be a few dB lower.

Histograms of vowel amplitudes for each sentence position are shown in figure 3-4. From this figure, it appears that most of the vowels have amplitudes between 0 and 25 dB below the overall peak. These data are relevant because amplitude (relative to overall peak) is an important cue for vowel landmarks (Mermelstein, in fact, requires that vowels be no more than 25 dB below the overall peak [61]).

3.5 Experiment 1: Presence of F1 Peak in Vowels

Prediction:

F1 will show a peak in a vowel between consonants (specifically, consonants with closures in the oral cavity) if the assumptions are accurate. The theoretical basis for this prediction is described in more detail in section 3.1.

Assumptions:

(a) The vowel is in fact an orally open vowel (not a syllabic nasal or lateral) between orally closed consonants (without elision, lenition, or other modifications). Exceptions are essentially failures of the database labeling to adequately represent the acoustic information (a syllabic sonorant not labeled as such, or a consonant not orally closed).

(b) F1 is clear and measurable, without interference by phenomena such as nasalization, glottalization, or aspiration. Also, the vowel is not reduced to the point where F1 is difficult to measure because of its short duration. Exceptions are likely to cause failure of the formant tracking algorithm.

(c) The vowel is produced with the oral cavity relatively open, and the consonants are produced with the oral cavity relatively closed. Also, the oral gesture is simply opening and closing, without complicated details such as lateralization. For amplitude measurements, the voicing source should be stable.

b d g p t k bcl dcl gcl pcl tcl kcl jh ch s sh z zh f th v dh m n ng em en eng	Closed
l r w y el iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr	Open
dx q nx hh hv ax-h pau epi h#	Other

Table 3.4: Experiment 1 classes for vowel contexts.

3.5.1 Methodology

Vowel tokens in the Raw experimental data (section 3.4.3) were classified for amplitude and frequency peak location (beginning, middle, or end). The classifications for amplitude and frequency were done separately, without referring to each other, because the expectation that amplitude and frequency peaks should coincide will be tested in Experiment 2 below.

The peak locations in time were quantized to 5.0 ms increments, because that was the frame period of the formant tracker. However, the aligned phonetic transcriptions were not quantized in time (floating point representation). Therefore, each peak was classified as “beginning” if it was within 5.0 ms of the vowel’s beginning, as labeled in the transcription, and “end” if it was within 5.0 ms of the vowel’s end. These data are referred to as the “Processed” data in following sections.

These Processed experimental data were analyzed for preceding and following context of vowels. For each vowel, the preceding and following tokens were classified as Closed oral cavity, Open oral cavity, or Other. The Closed class includes stops, fricatives, affricates, and nasals. The Open class includes vowels and semivowels. The Other class includes glottal stops (/q/), flaps (/dx nx/), aspirants (/hh hv/ and devoiced vowel /ax-h/), and silences. See Table 3.4 for the list.

3.5.2 Experiment 1A : Amplitude peaks against context

For Experiment 1A, context is compared to the location of the F1 amplitude peak. We begin with amplitude peaks, rather than frequency peaks, because we believe that amplitude peaks are likely to be more robust against error than frequency peaks (see discussion in section 3.4.2).

Preceding context may be C (closed), O (open), or X (other), and following context may also be C, O, or X. This gives a total of nine context categories. For each category, the amplitude peak may be at the beginning, middle, or end.

The theoretical prediction is that the amplitude peak should not appear at the beginning of the vowel in CV- contexts, and should not appear at the end of the vowel in -VC contexts. If the amplitude peak does appear in these unpredicted contexts, we expect that one or more of the underlying assumptions is not true.

Results

The basic statistics for the data set are shown in Table 3.5. Values representing unpredicted results are printed in emphatic typeface.

The data show that, in general, the theoretical prediction holds true. Of the CVC tokens in the database, almost all show an amplitude peak in the middle (27566 out of 27712, or 99.5%). When an amplitude peak appears at the beginning, it is almost always in OV-context (1622 out of 1639, or 99.0%), and when an amplitude peak appears at the end, it is almost always in -VO context (2636 out of 3015, or 87.4%).

Overall				
	C	O	X	All
C	27712 (34.3%)	9846 (12.2%)	5614 (6.94%)	43172 (53.4%)
O	16020 (19.8%)	5314 (6.57%)	3636 (4.50%)	24970 (30.8%)
X	8361 (10.3%)	2993 (3.70%)	1359 (1.68%)	12713 (15.7%)
All	52093 (64.4%)	18153 (22.5%)	10609 (13.1%)	80855 (100%)
Amplitude peak at beginning				
	C	O	X	All
C	<i>4 (0.00494%)</i>	0	0	<i>4 (0.00494%)</i>
O	1091 (1.35%)	166 (0.205%)	365 (0.451%)	1622 (2.00%)
X	12 (0.0148%)	1 (0.00124%)	0	13 (0.0161%)
All	1107 (1.37%)	167 (0.207%)	365 (0.451%)	1639 (2.03%)
Amplitude peak in middle				
	C	O	X	All
C	27566 (34.1%)	8252 (10.2%)	5467 (6.76%)	41285 (51.1%)
O	14917 (18.4%)	4541 (5.62%)	3249 (4.02%)	22707 (28.1%)
X	8323 (10.3%)	2557 (3.16%)	1329 (1.64%)	12209 (15.1%)
All	50806 (62.8%)	15350 (18.9%)	10045 (12.4%)	76201 (94.2%)
Amplitude peak at end				
	C	O	X	All
C	<i>142 (0.176%)</i>	1594 (1.97%)	147 (0.182%)	1883 (2.33%)
O	<i>12 (0.0148%)</i>	607 (0.751%)	22 (0.0272%)	641 (0.793%)
X	<i>26 (0.0322%)</i>	435 (0.538%)	30 (0.0371%)	491 (0.607%)
All	<i>180 (0.223%)</i>	2636 (3.26%)	199 (0.246%)	3015 (3.73%)

Table 3.5: Experiment 1A statistical results. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.

Fully 94.2% of all the vowels in the database show an amplitude peak in the middle. This supports the practice of using the amplitude peak as the primary indicator of vowel presence.

Anomalous or unexpected results

As theory predicts, very few of the amplitude peaks show up in unexpected contexts (184 tokens out of 80855, or 0.228%). For the peaks that do violate the theoretical prediction, each token was examined by hand for assumption violations. A small table of assumption violations was prepared (see Table 3.6, column 1) for this purpose. The table is organized to reflect the assumptions as described in section 3.5.

It should be pointed out that condition b3 “reduced” means that the vowel’s duration is so short that formant tracking is problematic, typically with only a few 5 millisecond estimates of formants. This indicates a problem which might be overcome by a different implementation of formant tracking, as opposed to the more theoretical reasons for formant tracking difficulty (nasalization, glottalization, or aspiration).

Each token was given a label from Table 3.6 if it appeared to violate the corresponding assumption, or two labels if it appeared to violate more than one. No token was given more than two labels, primarily because the labeling process is time consuming, and because adding more labels does not appear to provide significant new insights.

It was found that all of the tokens which produce unexpected results had at least one assumption violation, and many had two or three, or even more (see Table 3.6, column 3). Combinations of the B class (reasons why F1 is not clear or measurable) were especially frequent, with nasalization, glottalization, reduction, and aspiration occurring in combination. A typical example is the sequence / sh ix n /, frequently seen in words such as “motion,” in

Label	Meaning	Count
a1	Mislabeled	4
b1	Nasalized	114
b2	Glottalized	22
b3	Reduced	123
b4	Devoiced	84
c1	High vowel	2
c2	Source change	2
c3	Liquid context	10

Table 3.6: Experiment 1A assumption violations, labeled by hand.

which the vowel often displays several of these characteristics.

A total of 361 labels (from Table 3.6) were assigned to the 184 tokens which violated the theoretical prediction (averaging 1.96 per token). If more than two labels were assigned to each token, this average would probably be even higher.

To explore further, an automatic procedure was developed to look for evidence of some of these phenomena. The procedure uses only the Aligned Phonetic Transcription (APT) of the TIMIT database.

The automatic procedure is applied to each of the 184 anomalous tokens. It simply examines the segment label of the vowel, and the segment labels preceding and following the vowel, and looks for evidence that might indicate an assumption violation. Such evidence may be either in the vowel itself (for example, a syllabic nasal) or in its context (for example, a nasal segment following a vowel which has the F1 peak at the end).

The automatic procedure will not be nearly as accurate as hand labeling, because the APT does not capture all the nuances of vowel production. For instance, a vowel labeled as a schwa will probably be found to be more or less reduced, but whether the reduction is enough to justify a violation of the theoretical prediction is unclear. Conversely, many vowels which are

Label	Meaning	Count
a1	syllabic nasal	21
a2	consonant not closed	0
b1a	preceding nasal	13
b1b	following nasal	102
b2a	preceding glottal	18
b2b	following glottal	0
b3	schwa	154
b4a	devoiced	46
b4b	preceding aspirant	3
b4c	following aspirant	0
c1a	high	1
c1b	preceding high	1
c1c	following high	0
c2	voice source not stable	0
c3a	rhotic	1
c3b	preceding rhotic	4
c3c	following rhotic	0
c3d	syllabic lateral	2
c3e	preceding lateral	4
c3f	following lateral	0

Table 3.7: Experiment 1A assumption violations, labeled automatically.

heavily aspirated are not marked as such in the labeling. Therefore, the automatic procedure will be oversensitive to some phenomena and undersensitive to others. Still other phenomena are not captured by the automatic procedure at all. For instance, vowels in function words are more likely to be heavily reduced, and vowels at the end of the utterance are more likely to be glottalized.

The intent is to use the automatic procedure to winnow out those violations of theory which can be easily explained from context, and focus the hand labeler's attention on a small subset for further study.

A table of conditions was prepared, designed to follow the labels of Table 3.6 as much as possible. See Table 3.7, column 2.

The results are shown in Table 3.7, column 3. A total of 370 labels were assigned to the 184 tokens which violated the theoretical prediction (averaging 2.01 per token). Some of the results agree with current data on speech production. For instance, when nasal context is found, it is usually a following nasal, indicating that anticipatory nasalization of the vowel is more common and stronger than residual nasalization resulting from a preceding nasal consonant (see, for instance, [64]). Likewise, syllabic nasals usually have no other labels, indicating that this property alone is enough to confound the theoretical prediction, which makes sense because they are produced with oral closure (violating the first assumption in section 3.5).

Some other results are less compelling. For instance, many of the vowels are found to be schwas, but this does not mean that reduction is the principal mechanism at work, because there is no information about the length of these schwas.

Out of the 154 schwas found, 20 had no other labels attached, and these are the obvious candidates for further study. All showed an amplitude peak at the end of the vowel. Out of the 20 tokens labeled only as schwas, 10 were found to be shorter than 30 milliseconds in duration, short enough to be judged a sufficient reason for the peak location to be found close to a boundary (although many vowels shorter than 30 ms have peaks in the middle). The remaining 10 schwas were reexamined by hand and labeled for exception conditions, as shown in Table 3.8.

From the last column of Table 3.8 we can see that all the tokens in question have at least one characteristic that indicates an assumption violation. Interference with F1's measurability (the "B" category) is most common, primarily through glottalization and/or devoicing. There are two instances of mislabeling (the "A" category) and two instances of interference with oral opening (the "C" category).

Talker	Sent	#	ms	Ortho	Phonetic	Manual Label
fbcg1	sx442	11	30.4	thE gab	dh ix gcl	c2 function word pre stress
mctm0	si1350	16	33.7	-rams Of	z ax v	c2 function word pre stress
mdac2	sx369	5	39.2	ovEr the	v ax dh	a2 consonants very elided
mgak0	si1036	4	38.1	antithEsis	th ix s	b4a aspirated
mgsl0	sx354	10	44.9	exEcute	s ix k	a2 b4a consonants elided, aspirated
mjhi0	si1328	18	50.0	thE six	dh ax s	b2 heavily glottalized
mpgh0	sx114	9	46.4	and A tad-	nx ax tcl	b2 b4 glottalized and devoiced
mpgh0	sx384	4	40.9	-en thIs	dh ix s	b1 b4 nasalized and devoiced
mrfl0	si1156	2	36.6	this Is	s ix z	b4 devoiced
mses0	sx329	16	32.8	possIble	s ax v	a2 consonant heavily elided
medr0	sx384	4	35.8	-en thIs	dh ih s	b1 b4 nasalized and devoiced

Table 3.8: Experiment 1A residual assumption violations, labeled by hand. The Talker and Sentence identify the utterance, and the Index identifies the vowel in question, numbered sequentially from the beginning. The vowel length is in milliseconds. The vowel in question is capitalized in the Orthographic fragment.

Only one token (medr0 sx384 4) received no automatic label at all – it appears on the last line of table 3.8. The word is “this” with a high front vowel, which the phonetic transcription marks as lax but not reduced (even though it is only 35 ms long). The vowel appears rather nasalized (from the preceding word “shorten”) and devoiced (from the following fricative cluster in “this skirt”).

It is evident from Table 3.7 that nasalization (especially from following nasals) is a relatively frequent problem. Acoustic theory predicts that nasalization will often cause reduction of F1 amplitude. Ultimately, the first pass of a Vowel Landmark Detector will probably need to be followed by a second pass, after nasal context has been detected by other means. See section 5.3.2 for more discussion of error recovery.

Conclusions to Experiment 1A

Overall, the experimental results agree very well with the theoretical prediction. Over 94% of all vowel tokens in the database show an amplitude peak somewhere in the middle of the vowel, which is the most important finding for a Vowel landmark detector.

Of the vowel tokens which do not show an amplitude peak in the middle, all were found to have one or more violations of the assumptions which underlie the theoretical prediction. Most of these conditions were detectable by a simple automatic procedure which examines the token's context and duration, and the rest were detectable by manual inspection.

3.5.3 Experiment 1B: Frequency peaks against context

For Experiment 1B, context is compared to the location of the F1 frequency peak. As described above (section 3.5.2), there is reason to expect that the frequency peaks will be less well behaved than the amplitude peaks, because of errors from the formant tracker.

The database, context categories (as shown in table 3.4), and processing are the same as for Experiment 1A, but applied to the frequency peaks instead of the amplitude peaks.

Results

The basic statistics for the data set are shown in Table 3.9. Values representing unpredicted results are printed in emphatic typeface.

Overall				
	C	O	X	All
C	27712 (34.3%)	9846 (12.2%)	5614 (6.94%)	43172 (53.4%)
O	16020 (19.8%)	5314 (6.57%)	3636 (4.50%)	24970 (30.9%)
X	8361 (10.3%)	2993 (3.70%)	1359 (1.68%)	12713 (15.7%)
All	52093 (64.4%)	18153 (22.5%)	10609 (13.1%)	80855 (100%)
Frequency peak at beginning				
	C	O	X	All
C	<i>35 (0.0433%)</i>	<i>11 (0.0136%)</i>	<i>2 (0.00247%)</i>	<i>48 (0.0594%)</i>
O	1350 (1.67%)	211 (0.261%)	173 (0.214%)	1734 (2.14%)
X	142 (0.176%)	53 (0.0655%)	10 (0.0124%)	205 (0.254%)
All	1527 (1.89%)	275 (0.340%)	185 (0.229%)	1987 (2.46%)
Frequency peak in middle				
	C	O	X	All
C	26734 (33.1%)	7800 (9.65%)	4915 (6.08%)	39449 (48.8%)
O	14187 (17.5%)	3705 (4.58%)	3168 (3.92%)	21060 (26.0%)
X	7984 (9.87%)	2448 (3.03%)	1100 (1.36%)	11532 (14.3%)
All	48905 (60.5%)	13953 (17.3%)	9183 (11.4%)	72041 (89.1%)
Frequency peak at end				
	C	O	X	All
C	<i>943 (1.17%)</i>	2035 (2.52%)	697 (0.862%)	3675 (4.55%)
O	<i>483 (0.597%)</i>	1398 (1.73%)	295 (0.365%)	2176 (2.69%)
X	<i>235 (0.291%)</i>	492 (0.608%)	249 (0.308%)	976 (1.21%)
All	<i>1661 (2.05%)</i>	3925 (4.85%)	1241 (1.53%)	6827 (8.44%)

Table 3.9: Experiment 1B statistical results. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.

The data show that the theoretical prediction holds true, for the most part. The results of this experiment show roughly an order of magnitude more unexpected peaks than the amplitude experiment did. Of all the tokens in the database, most show a frequency peak in the middle (89.1%). Of the CVC tokens in the database, almost all show a frequency peak in the middle (26734 out of 27712, or 96.5%). When a frequency peak appears at the beginning, it is usually in OV- context (1734 out of 1987, or 87.3%), and when a frequency peak appears at the end, it is often in -VO context (3925 out of 6827, or 57.7%). However, these data are far more ambiguous than the results of the amplitude experiment.

Presumably, most of these unexpected results come from errors in the formant tracker. One characteristic of most formant trackers is increased error rates for higher pitched voices, whose wider spacing of harmonics makes the formants more difficult to track. To see whether this is a factor in the current experiment, we separate the results by talker gender. If widely spaced harmonics are contributing to formant tracker error, we expect to see female voices show higher error rates.

Results for female talkers only (24725 tokens) are shown in table 3.10, and results for male talkers only (56131 tokens) are shown in table 3.11. There is not much evidence for a systematic effect of talker gender on these statistics. Most of the percentages are very close, and the predicted data for males are not consistently better than for females (which would be the case if higher pitch were a major cause of formant tracker errors). Thus, we conclude that higher pitch alone is not a major contribution to errors of the formant tracker.

Another possible source of error is boundary phenomena. The formant tracker is probably more likely to make errors at the edges of vowels, where the influence of adjacent consonants (nasalization, frication, and so forth) is greatest. Such problems are especially likely if the segment labels are not placed accurately in time, which has been observed as a persistent problem with the TIMIT database.

Overall				
	C	O	X	All
C	8616 (34.8%)	2935 (11.9%)	1689 (6.83%)	13240 (53.6%)
O	4890 (19.8%)	1551 (6.27%)	1081 (4.37%)	7522 (30.4%)
X	2633 (10.6%)	896 (3.62%)	433 (1.75%)	3962 (16.0%)
All	16139 (65.3%)	5382 (21.8%)	3203 (12.9%)	24724 (100%)
Frequency peak at beginning				
	C	O	X	All
C	<i>11 (0.0445%)</i>	<i>4 (0.0162%)</i>	<i>1 (0.00404%)</i>	<i>16 (0.0647%)</i>
O	420 (1.70%)	36 (0.146%)	52 (0.210%)	508 (2.05%)
X	45 (0.182%)	14 (0.0566%)	2 (0.00809%)	61 (0.247%)
All	476 (1.93%)	54 (0.218%)	55 (0.222%)	585 (2.37%)
Frequency peak in middle				
	C	O	X	All
C	8344 (33.7%)	2376 (9.61%)	1487 (6.01%)	12207 (49.4%)
O	4349 (17.6%)	1119 (4.53%)	939 (3.80%)	6407 (25.9%)
X	2529 (10.2%)	764 (3.09%)	333 (1.35%)	3626 (14.7%)
All	15222 (61.6%)	4259 (17.2%)	2759 (11.2%)	22240 (89.9%)
Frequency peak at end				
	C	O	X	All
C	<i>261 (1.06%)</i>	555 (2.24%)	201 (0.813%)	1017 (4.11%)
O	<i>121 (0.489%)</i>	396 (1.60%)	90 (0.364%)	607 (2.46%)
X	<i>59 (0.239%)</i>	118 (0.477%)	98 (0.396%)	275 (1.11%)
All	<i>441 (1.78%)</i>	1069 (4.32%)	389 (1.57%)	1899 (7.68%)

Table 3.10: Experiment 1B statistical results, female talkers only

Overall				
	C	O	X	All
C	19096 (34.0%)	6911 (12.3%)	3925 (6.99%)	29932 (53.3%)
O	11130 (19.8%)	3763 (6.70%)	2555 (4.55%)	17448 (31.1%)
X	5728 (10.2%)	2097 (3.74%)	926 (1.65%)	8751 (15.6%)
All	35954 (64.1%)	12771 (22.8%)	7406 (13.2%)	56131 (100%)
Frequency peak at beginning				
	C	O	X	All
C	24 (0.0428%)	7 (0.0125%)	1 (0.00178%)	32 (0.0570%)
O	930 (1.66%)	175 (0.312%)	121 (0.216%)	1226 (2.18%)
X	97 (0.173%)	39 (0.0695%)	8 (0.0143%)	144 (0.257%)
All	1051 (1.87%)	221 (0.394%)	130 (0.237%)	1402 (2.50%)
Frequency peak in middle				
	C	O	X	All
C	18390 (32.8%)	5424 (9.66%)	3428 (6.11%)	27242 (48.5%)
O	9838 (17.5%)	2586 (4.61%)	2229 (3.97%)	14653 (26.1%)
X	5455 (9.72%)	1684 (3.00%)	767 (1.37%)	7906 (14.1%)
All	33683 (60.0%)	9694 (17.3%)	6424 (11.4%)	49801 (88.7%)
Frequency peak at end				
	C	O	X	All
C	682 (1.22%)	1480 (2.64%)	496 (0.884%)	2658 (4.74%)
O	362 (0.645%)	1002 (1.79%)	205 (0.365%)	1569 (2.80%)
X	176 (0.314%)	374 (0.666%)	151 (0.269%)	701 (1.25%)
All	1220 (2.17%)	2856 (5.09%)	852 (1.52%)	4928 (8.78%)

Table 3.11: Experiment 1B statistical results, male talkers only

iy ih uh uw	High
eh ey ao ow er	Mid
ae aa ah	Low

Table 3.12: Vowel height classes. Schwas, diphthongs, and syllabic sonorants are not included, because their acoustic manifestation of height is liable to be uncertain or ambiguous.

Vowel Height	F1 Peak Location		
	Begin	Middle	End
Token Counts			
High	250	14668	2986
Mid	392	16695	1149
Low	93	12421	272
Mean Frequency (Hz)			
High	483	478	535
Mid	564	607	655
Low	649	702	861

Table 3.13: Experiment 1B statistical results, by vowel height

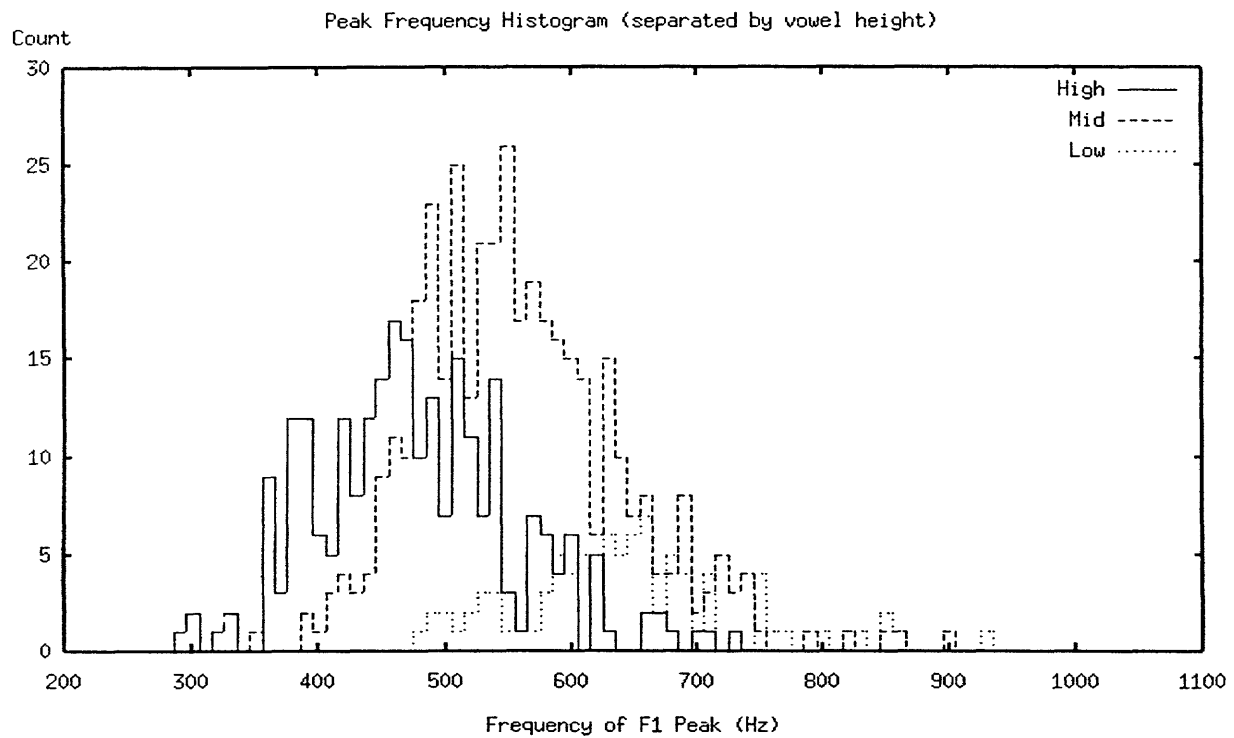


Figure 3-5: Beginning Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak at the beginning, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.

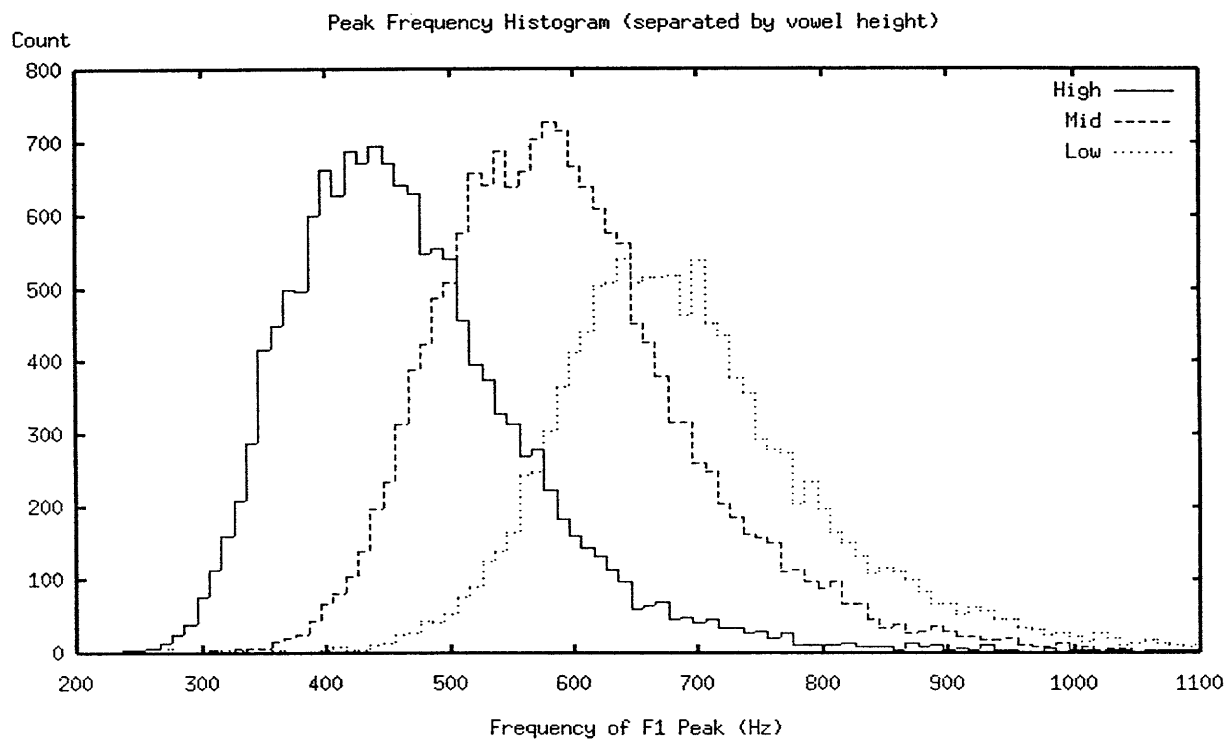


Figure 3-6: Middle Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak in the middle, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.

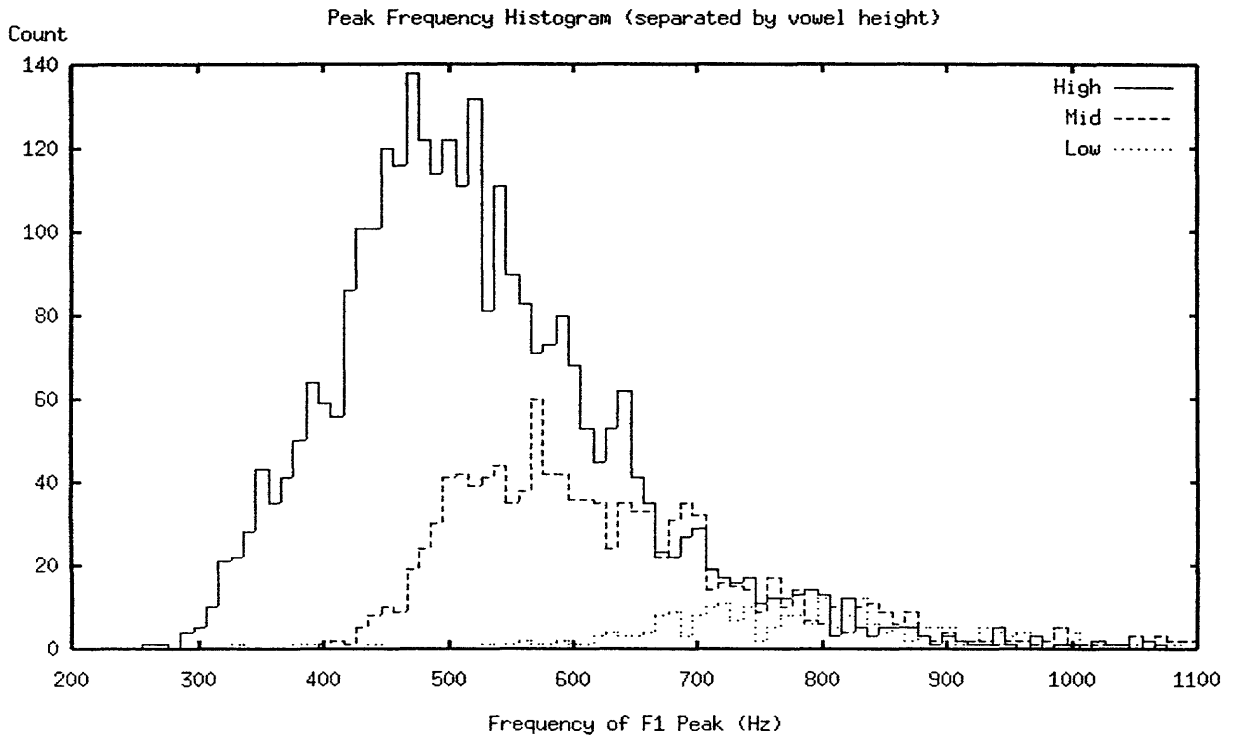


Figure 3-7: End Peak Frequency Histogram. This plot shows the frequency of the peak of the F1 track, for vowels with the peak at the end, separated by vowel height. Some few tokens are outside the frequency bounds of this histogram.

An erroneously high F1 measurement at the boundary is likely to be detected as the F1 peak. (An erroneously low measurement will not affect the data in this experiment.) If this phenomenon is present in the data, we expect the peak frequency values which are detected at the endpoints to be higher (on average) than the peak values detected in the middle of the vowel. Also, we expect the peak values in the middle to correlate well with the vowel height, while the peak values at the endpoints should not correlate as well with vowel height.

Vowel height was assigned to the TIMIT labels as shown in table 3.12.³ For this experiment, vowels whose height is liable to be uncertain or ambiguous (schwas, diphthongs, and syllabic sonorants) are excluded, and only the tabulated vowels are used.

Basic statistics for the peak frequency values (separated by peak location and vowel height) are shown in table 3.13. The peaks located at the end of the vowels are significantly higher in frequency than the peaks located in the middle of the vowels, for all vowel heights. (The peaks located at the beginning of vowels do not show a consistent pattern, but they are much less numerous.) This indicates erroneous formant estimates at the vowel boundaries (at least at the ending boundaries). Perhaps postvocalic segments (which are liable to be weaker than prevocalic segments) tend to be labeled erroneously.

Table 3.13 also shows that peaks at the end of vowels occur much more frequently for high vowels than for mid vowels, and much more frequently for mid vowels than for low vowels. This makes sense because F1 is intrinsically lowest in high vowels, and therefore it's more likely to see it peak at the boundary.

The mean frequency of the peaks does correlate fairly well with vowel height, for all peak locations. There does not seem to be a clear difference between the different peak locations. Histograms of peak frequency for the different peak locations are shown in figures 3-5 (begin-

³The vowel /ah/ is sometimes considered non low.

ning), 3-6 (middle), and 3-7 (end). Figure 3-6 does show a clear correlation between vowel height and frequency for middle peaks, as expected, and the frequency values are in good agreement with previous studies of how F1 frequency varies with vowel height (for example, [66, p. 183]). Figure 3-7 shows (again) that end peaks are more frequent for higher vowels, but there is still correlation between vowel height and frequency. Figure 3-5 does not show much consistency, probably because of the sparseness of the data.

All of these figures show rather more spreading of the peaks than some other studies [18], which may be partly due to the data including both male and female talkers. Female talkers tend to have a slightly shorter vocal tract, and therefore slightly higher average formant values. Genders are not separated in this experiment, because our goal is a vowel landmark detector that is gender independent. Also, the TIMIT labelling may not adequately represent phenomena such as reduction and assimilation, which will also cause more variation in the F1 frequency.

As in Experiment 1A, there are substantially fewer peaks at the beginning than at the end. Syllable-initial consonants, especially when prestressed or word initial [30], are acoustically clearer and phonologically more robust than syllable-final consonants, and therefore more likely to contrast with the following vowel.

In summary, there is some evidence for erroneous formant estimates at the vowel boundaries, but it is not conclusive. To explore further, the peak picking search was redone, with the vowel boundaries moved inward from the time points given in the TIMIT transcription. The vowel boundaries were moved inward 10 milliseconds from each end. Some vowels (about 850) are short enough that the resulting duration is too short for the formant tracker to operate accurately, and these were discarded before the statistical analysis.

Results of this experiment are shown in table 3.14. In comparison with table 3.9, the

Overall				
	C	O	X	All
C	27164 (33.9%)	9798 (12.2%)	5549 (6.94%)	42511 (53.1%)
O	15929 (19.9%)	5300 (6.62%)	3627 (4.53%)	24856 (31.1%)
X	8295 (10.4%)	2984 (3.73%)	1356 (1.69%)	12635 (15.8%)
All	51388 (64.2%)	18082 (22.6%)	10532 (13.2%)	80002 (100%)
Frequency peak at beginning				
	C	O	X	All
C	<i>7567 (9.46%)</i>	<i>1453 (1.82%)</i>	<i>1137 (1.42%)</i>	<i>10157 (12.7%)</i>
O	6243 (7.80%)	1315 (1.64%)	883 (1.10%)	8441 (10.6%)
X	2664 (3.33%)	754 (0.942%)	228 (0.285%)	3646 (4.56%)
All	16474 (20.6%)	3522 (4.40%)	2248 (2.81%)	22244 (27.8%)
Frequency peak in middle				
	C	O	X	All
C	17192 (21.5%)	5715 (7.14%)	3407 (4.26%)	26314 (32.9%)
O	8666 (10.8%)	2331 (2.91%)	2259 (2.82%)	13256 (16.6%)
X	5066 (6.33%)	1612 (2.01%)	786 (0.982%)	7464 (9.33%)
All	30924 (38.7%)	9658 (12.1%)	6452 (8.06%)	47034 (58.8%)
Frequency peak at end				
	C	O	X	All
C	<i>2405 (3.01%)</i>	2630 (3.29%)	1005 (1.26%)	6040 (7.55%)
O	<i>1020 (1.27%)</i>	1654 (2.07%)	485 (0.606%)	3159 (3.95%)
X	<i>565 (0.706%)</i>	618 (0.772%)	342 (0.427%)	1525 (1.91%)
All	<i>3990 (4.99%)</i>	4902 (6.13%)	1832 (2.29%)	10724 (13.4%)

Table 3.14: Experiment 1B statistical results for truncated vowels. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.

truncated vowels show substantially more tokens which violate theoretical predictions. Also, the truncated vowels show substantially more peaks at the beginnings of vowels than at the ends, in direct contrast to table 3.9. This is (at least) an indication that very different phenomena are at work in the truncated vowel data set, which is some evidence for erroneous formant estimates at the vowel boundaries (but what phenomena is not clear).

Presumably, many (if not most) of the vowel tokens which do not show a peak in the middle should have predictable characteristics. High vowels are liable to have lower F1 frequency

	Frequency peak at beginning						
	high	mid	low	schwa	sonorant	diphth	All
stop	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
fric	0 (0)	0 (0)	0 (0)	1 (0.05)	0 (0)	0 (0)	1 (0.05)
affric	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
nasal	7 (0.35)	4 (0.20)	3 (0.15)	28 (1.41)	5 (0.25)	0 (0)	47 (2.37)
semi	30 (1.51)	9 (0.45)	5 (0.25)	46 (2.32)	0 (0)	2 (0.10)	92 (4.6)
vowel	186 (9.36)	335 (16.9)	49 (2.47)	1027 (51.7)	35 (1.76)	10 (0.50)	1642 (82.6)
other	27 (1.36)	44 (2.21)	36 (1.81)	86 (4.33)	1 (0.05)	11 (0.55)	205 (10.3)
All	250 (12.6)	392 (19.7)	93 (4.68)	1188 (59.8)	41 (2.06)	23 (1.16)	1987

Table 3.15: Experiment 1B context statistics, for frequency peaks at beginning of segment. Columns show the class of the vowel token, and rows show the manner of the preceding segment. Each entry shows the token count followed by the percent in parentheses.

than an adjacent semivowel, for instance. Such phenomena should appear in a study of all vowel tokens which do not show a peak in the middle.

1987 out of the 80856 vowel tokens (2.46%) show a frequency peak at the beginning of the labeled segment. Table 3.15 shows the statistics of these tokens, by type of vowel and by manner of the preceding segment.

Over 80% of these tokens are preceded by vowels (over 50% are schwas preceded by vowels). Of the remainder, most are preceded by segments with manner "other" (which includes pauses and sentence boundaries). Semivowels are the most frequent context other than vowels. High and mid vowels are more frequent than low vowels.

6827 out of the 80856 vowel tokens (8.44%) show a frequency peak at the end of the labeled segment. Table 3.16 shows the statistics of these tokens, by type of vowel and by manner of the preceding segment.

High vowels (43.7%) and schwas (28.0%) account for over two-thirds of these tokens. Most occur in the context of other vowels (40.7%) and semivowels (21.4%). High vowels followed

	Frequency peak at end						
	high	mid	low	schwa	sonorant	diphth	All
stop	246 (3.60)	169 (2.48)	70 (1.03)	436 (6.39)	91 (1.33)	19 (0.28)	1031 (15.1)
fric	270 (3.95)	130 (1.90)	112 (1.64)	334 (4.89)	55 (0.81)	33 (0.48)	934 (13.7)
affric	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
nasal	44 (0.64)	29 (0.42)	20 (0.29)	298 (4.37)	3 (0.04)	6 (0.09)	400 (5.86)
semi	626 (9.17)	417 (6.11)	60 (0.88)	297 (4.35)	49 (0.72)	12 (0.18)	1461 (21.4)
vowel	1747 (25.6)	378 (5.54)	4 (0.06)	442 (6.47)	192 (2.81)	13 (0.19)	2776 (40.7)
other	53 (0.78)	26 (0.38)	6 (0.09)	103 (1.51)	31 (0.45)	6 (0.09)	225 (3.30)
All	2986 (43.7)	1149 (16.8)	272 (3.98)	1910 (28.0)	421 (6.17)	89 (1.30)	6827

Table 3.16: Experiment 1B context statistics, for frequency peaks at end of segment. Columns show the class of the vowel token, and rows show the manner of the following segment. Each entry shows the token count followed by the percent in parentheses.

by either vowels or semivowels are by far the most frequent examples, confirming the general prediction.

Anomalous or unexpected results

Even though the frequency peak data show an order of magnitude more errors than the amplitude data (section 3.5.2), it is still true that very few of the frequency peaks appear in unexpected contexts (1709 out of 80855, or 2.11%). Those peaks that do violate the theoretical prediction were examined with the same automatic procedure that was used on the amplitude data (as shown in Table 3.7).

The results are shown in Table 3.17. A total of 2182 violation labels were assigned to the 1709 tokens (averaging 1.28 per token).

Of the 326 tokens which received no label at all, only 2 are shorter than 30 milliseconds. Of the 226 tokens which received only the label b3 (schwa), only 41 are shorter than 30 milliseconds. The remaining tokens (324 unlabeled and 185 long schwas) total 509, which is

Label	Meaning	Count
a1	syllabic nasal	82
b1a	preceding nasal	207
b1b	following nasal	347
b2a	preceding glottal	82
b2b	following glottal	1
b3	schwa	704
b4a	devoiced	52
b4b	preceding aspirant	38
b4c	following aspirant	0
c1a	high	196
c1b	preceding high	39
c1c	following high	0
c3a	rhotic	103
c3b	preceding rhotic	189
c3c	following rhotic	5
c3d	syllabic lateral	33
c3e	preceding lateral	99
c3f	following lateral	5

Table 3.17: Experiment 1B assumption violations, labeled automatically.

a substantially higher number of residual anomalies than in the amplitude data.

To examine all 509 of these residuals by hand would be tedious and time consuming, and only limited insight would be gained. Hand examination was judged to be unnecessary in this case.

Conclusions to Experiment 1B

Overall, the experimental results agree fairly well with the theoretical prediction. Over 89% of all vowel tokens in the database show a frequency peak somewhere in the middle of the vowel.

3.5.4 Conclusions to Experiment 1

The experimental results agree with the theoretical prediction. The vast majority of vowel tokens show peaks (in both frequency and amplitude) somewhere in the middle of the vowel. Amplitude peaks are more consistent than frequency peaks.

3.6 Experiment 2: Coincidence of Amplitude and Frequency Peaks in Vowels

Prediction:

schwa	20470
sonorant	1908
lax	19235
tense	23361
diphthong	4980
All	69954

Table 3.18: Experiment 2 vowel categories and counts. For each category, these data are counts of all vowels with amplitude and frequency peaks both in the middle of the vowel.

F1's amplitude peak and frequency peak will occur at the same place in time.

Assumptions:

(a) F1 has in fact a definite peak, not at the endpoints.

(b) F1 is clear and measurable, without interference by phenomena such as nasalization, glottalization, extreme reduction, or aspiration. Exceptions are likely to cause failure of the formant tracking algorithm.

(c) The voicing source is stable (not changing in amplitude or in spectral shape), nasalization is not affecting the F1 peak, and higher formants (particularly F2) are not extremely close to F1. Exceptions are likely to affect the amplitude peak, so as to make it appear in a different place from the frequency peak.

3.6.1 Methodology

The Processed experimental data (see section 3.5.1) were filtered to keep only those tokens which have amplitude and frequency peaks both in the middle of the vowel. Table 3.18 shows the counts of vowel tokens in each category.

For each vowel, the peak location was computed as a percentage of the vowel's duration. In this way, the peak locations were normalized for duration, and histograms of peak location in normalized duration were plotted (figures 3-8 and 3-9).

The peak times are quantized to 5 millisecond intervals (because the formant tracker outputs a value every 5 ms). Such quantization causes an artifact in the histogram, whereby short vowels cause a preponderance of counts in the bins which are low divisors of the normalized duration ($1/2$ and $1/4$ of duration are most evident, $1/3$ and $2/3$ less evident but still visible). The resulting artifacts might be called "crenellation," since they look somewhat like the ramparts of a medieval castle.

In order to avoid crenellation artifacts and produce a smoother appearing histogram, a "dithering" algorithm was used. When a vowel is short enough that the 5 ms quantization value spans N histogram bins, each of N bins (centered on the peak location) has its count increased by $1/N$. This technique avoids crenellation and yields a smooth histogram, as seen in figures 3-8 and 3-9.

Results

The data in figures 3-8 and 3-9 clearly show that peaks (both frequency and amplitude) tend to appear early in the vowel rather than late (although there are examples of peaks at all times). This phenomenon is most apparent for tense vowels, and least apparent for lax vowels. Tense vowels (at least those which are not +low) tend to offglide towards /i/ or /u/, i. e. towards a low F1, while lax vowels do not.

The data also show that the frequency peaks tend to appear earlier than the amplitude peaks. This phenomenon is in contrast to the theoretical prediction. To examine it further, a joint two-dimensional histogram of frequency and amplitude peaks was created, and plotted in

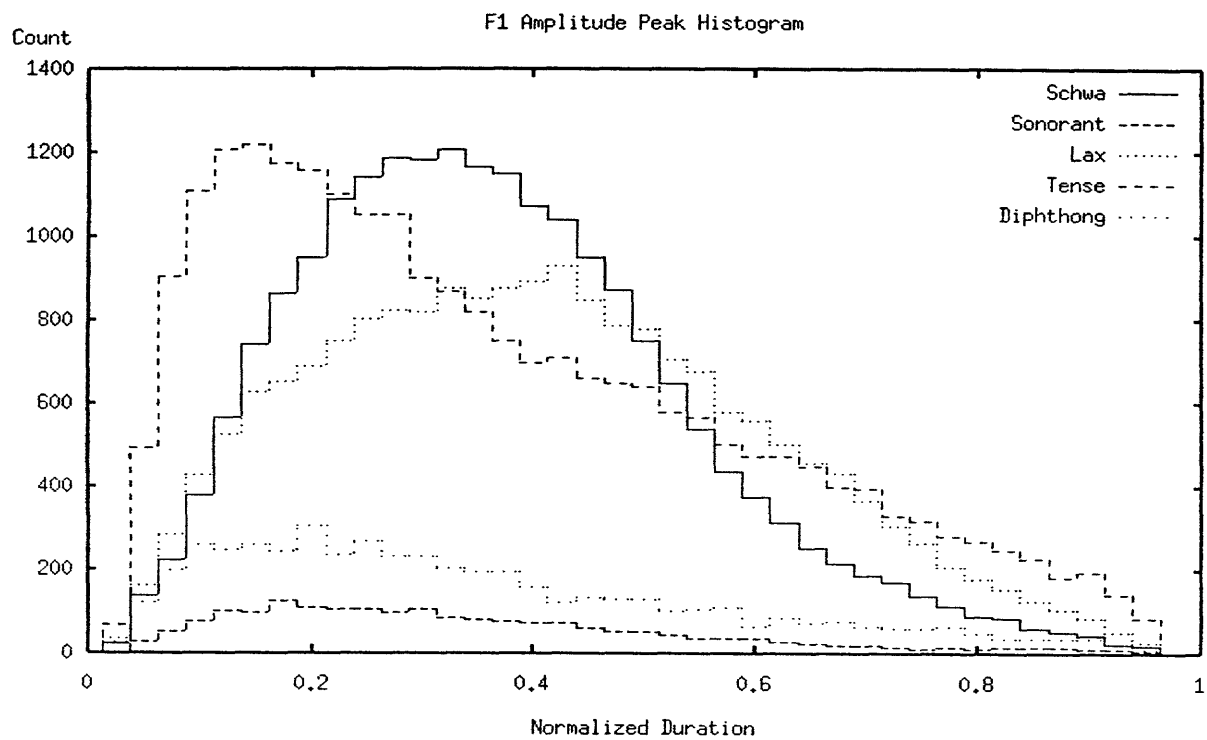


Figure 3-8: F1 Amplitude Peak Histogram. The data include all vowels in table 3.18. The horizontal axis is the amplitude peak location, normalized against the duration of the vowel. The histogram is dithered to avoid crenellation artifacts.

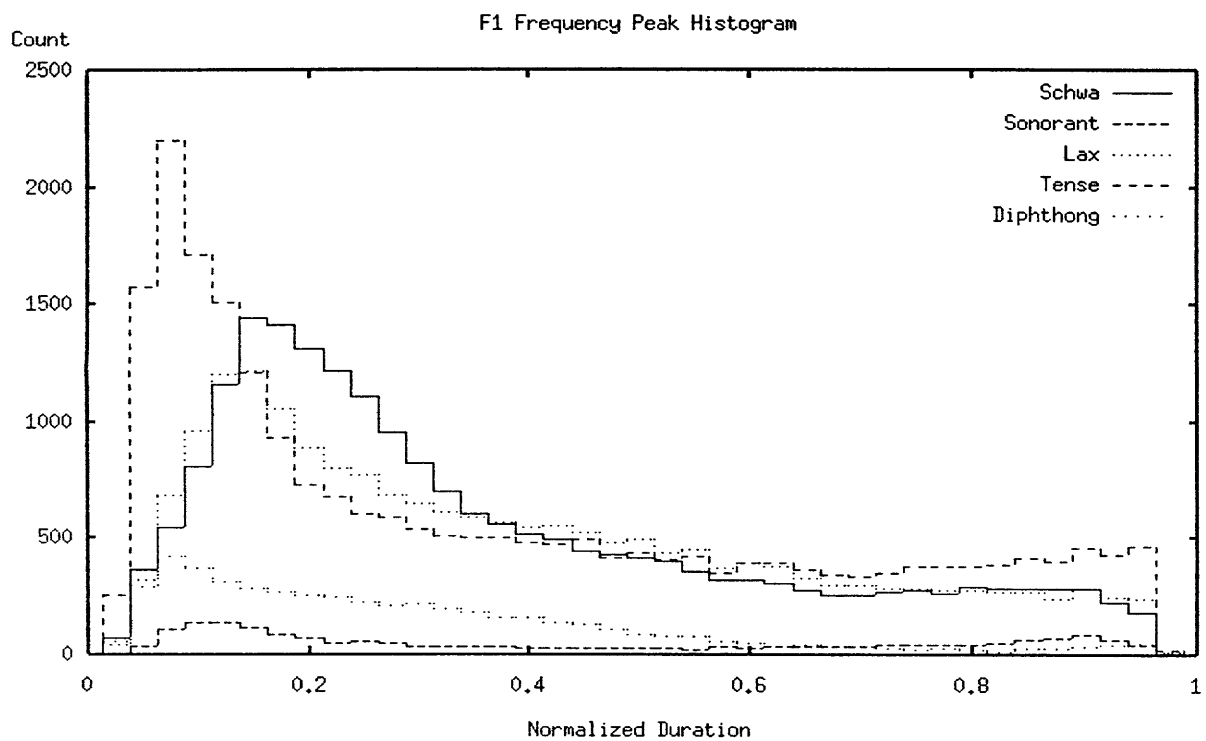


Figure 3-9: F1 Frequency Peak Histogram. The data include all vowels in table 3.18. The horizontal axis is the frequency peak location, normalized against the duration of the vowel. The histogram is dithered to avoid crenellation artifacts.

F1 Peak Cross Histogram

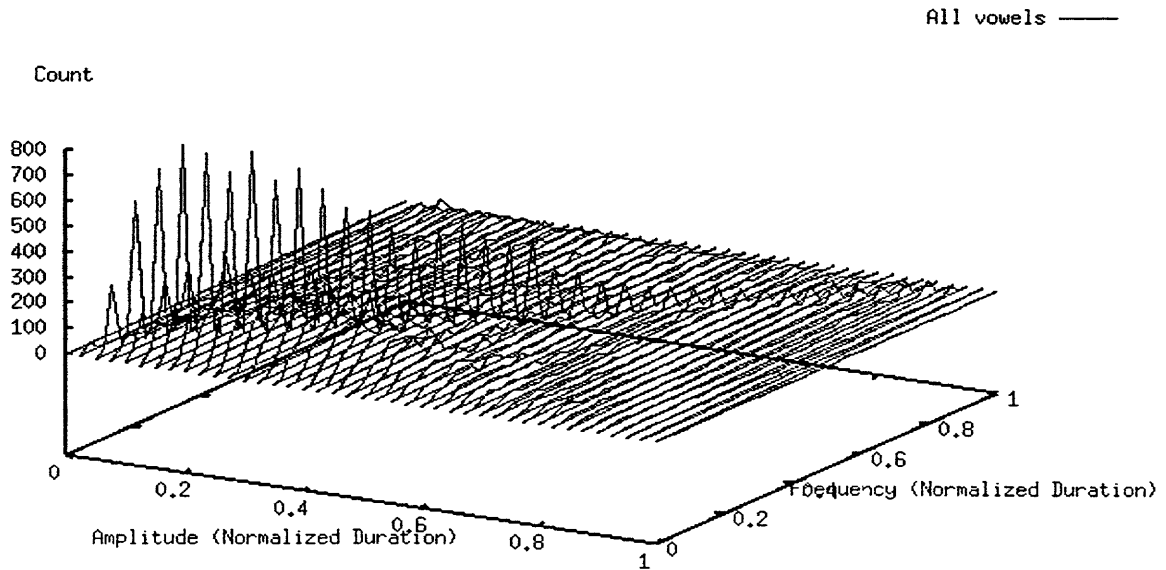


Figure 3-10: F1 Amplitude and Frequency Peak Cross Histogram. The data include all vowels in table 3.18. The horizontal axes are amplitude and frequency peak locations, respectively, normalized against the duration of the vowel.

a three dimensional projection as shown in figure 3-10. In this depiction, the theoretical prediction is that all points should lie along the diagonal.

The cross histogram clearly shows a preponderance of tokens along the diagonal (as predicted by theory), especially with peaks towards the beginning of the vowel. It also shows some tokens off the diagonal (in contrast to the theoretical prediction), but this projection does not clearly show what patterns may be evident in the off-diagonal tokens. Several attempts to make this projection clearer (by rotation and scaling) were not very successful.

In an effort to see the patterns more clearly, a histogram of tokens' separation from the

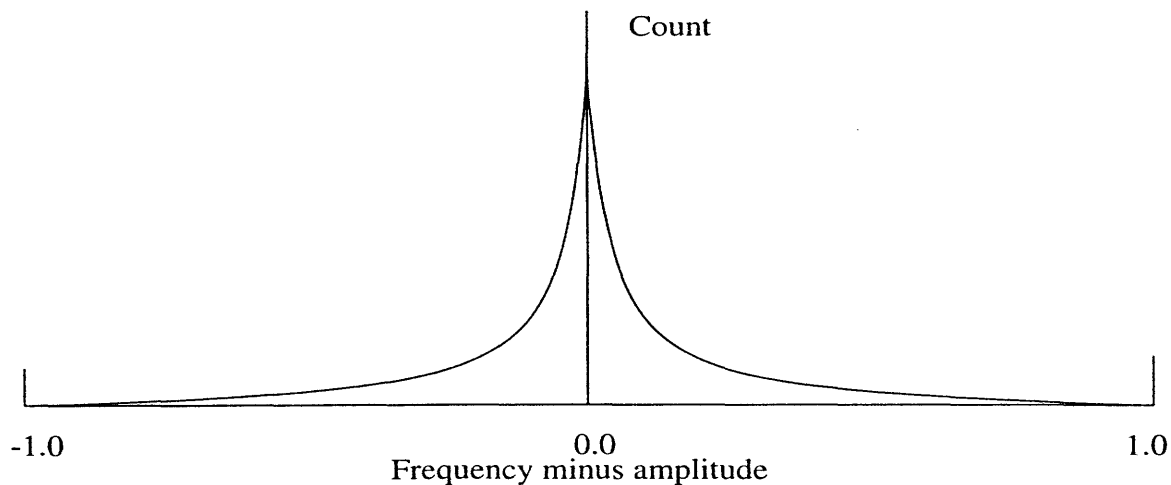


Figure 3-11: Schematic of expected histogram of F1 peaks on the difference diagonal.

diagonal was computed. In essence, this consists of rotating the histogram of figure 3-10 by 45 degrees, to place the diagonal along one axis, and summing the histogram bins along the other axis. Such a rotation is properly performed by multiplication by a rotation matrix of the form

$$\begin{vmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{vmatrix} \quad (3.1)$$

However, for a rotation angle of 45 degrees, $\cos \theta = \sin \theta$ and a simple difference of frequency and amplitude is sufficient. (There is a scale factor of 0.7071 which may be included, but merely changes the histogram's horizontal axis.) We expect the resulting "difference diagonal" histogram to have a high peak at zero, and taper off rapidly on both sides, as shown schematically in Figure 3-11.

The "difference diagonal" histogram was computed from the same data set as the two dimen-

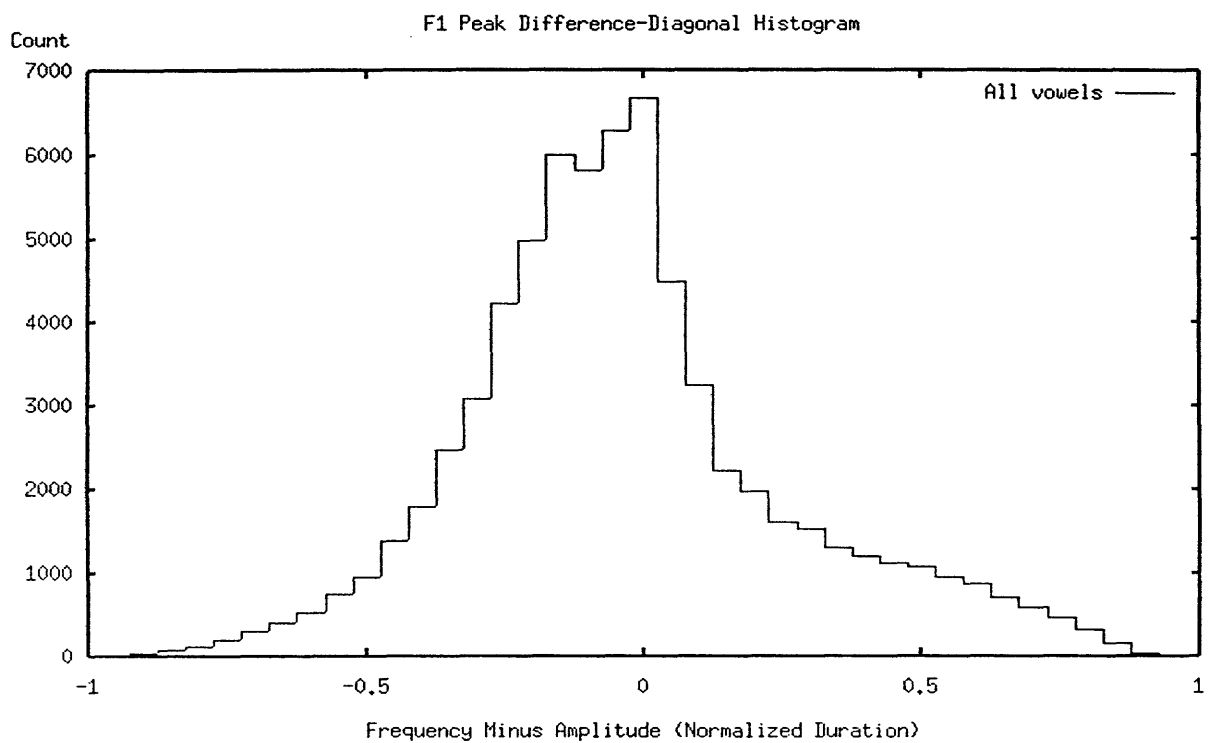


Figure 3-12: F1 Peak Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin.

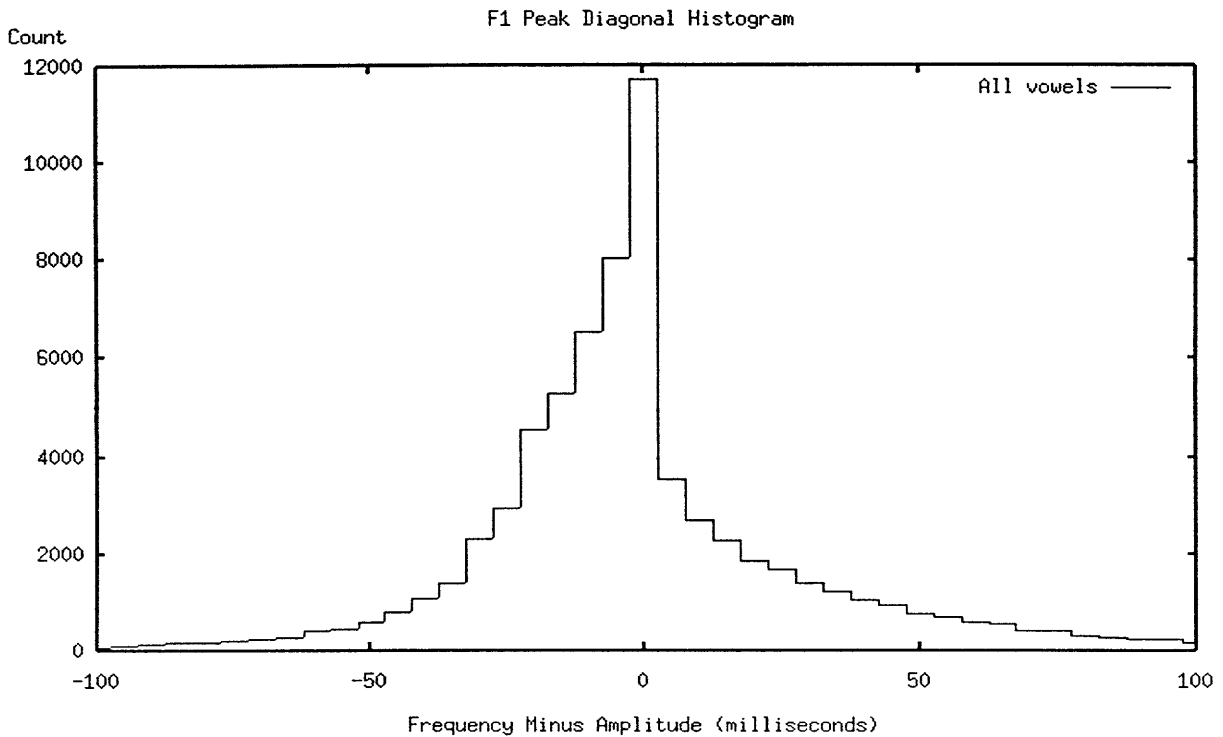


Figure 3-13: F1 Peak Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, in milliseconds. The vertical axis is the number of tokens per bin.

sional histogram, and the result is shown in figure 3-12. The basic prediction is validated, as the data show a high peak on the diagonal, and taper off rapidly for tokens off the diagonal. However, the figure also shows a pronounced “shoulder” just below zero on the horizontal axis, indicating a tendency for frequency peaks to occur earlier than amplitude peaks.

Figure 3-13 shows the same data, plotted as a function of the literal difference (frequency minus amplitude in milliseconds). It shows the same general pattern as Figure 3-12, with the “shoulder” for frequency peaks occurring earlier than amplitude peaks, but appears more concentrated around the zero point. By comparison, Figure 3-12 appears more spread out, because the normalized duration emphasizes the difference in short segments.

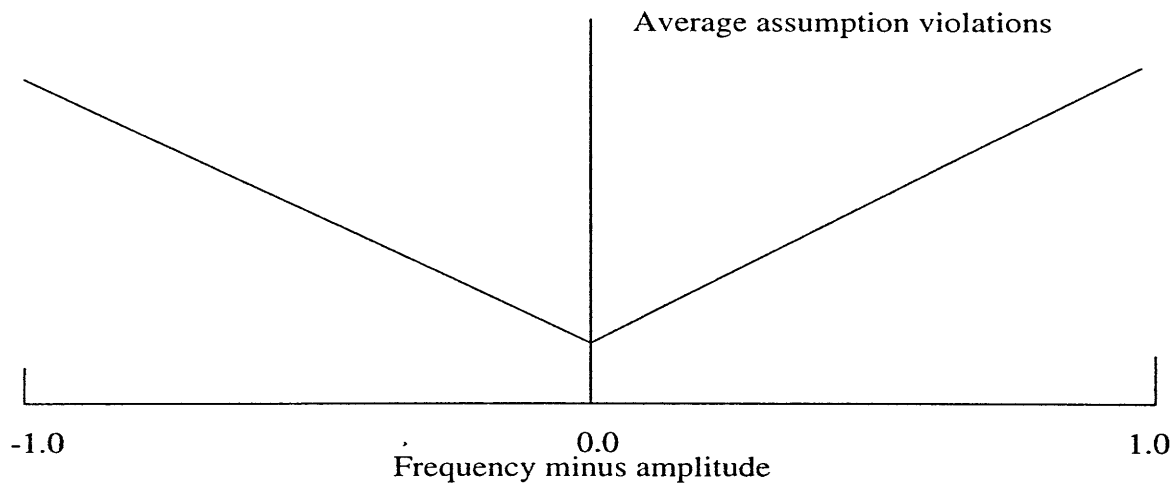


Figure 3-14: Schematic of expected assumption violations among F1 peaks on the difference diagonal.

Although most of the tokens are close to the diagonal in figure 3-12, there are a significant number off the diagonal, in violation of the theoretical prediction. It may be surmised that these anomalous tokens should show violations of the assumptions underlying the theoretical prediction, and that there should be more assumption violations (or more extreme violations) for tokens very far from the diagonal. If a histogram of assumption violations per token were plotted in the same way as figure 3-12, it should show a minimum at zero, and rising values towards both positive and negative extremes, as shown schematically in Figure 3-14.

Violations of theoretical assumptions were detected using the same procedure as in section 3.5.2, and tabulated as in Table 3.7. Using these data, the assumption violation histogram was computed from the same data set as the diagonal histogram, and the result is shown in figure 3-15. Again, the basic prediction is validated, as the data show rising values towards both positive and negative extremes.

However, there is a noticeable bump at the zero point, which is not part of the prediction. It

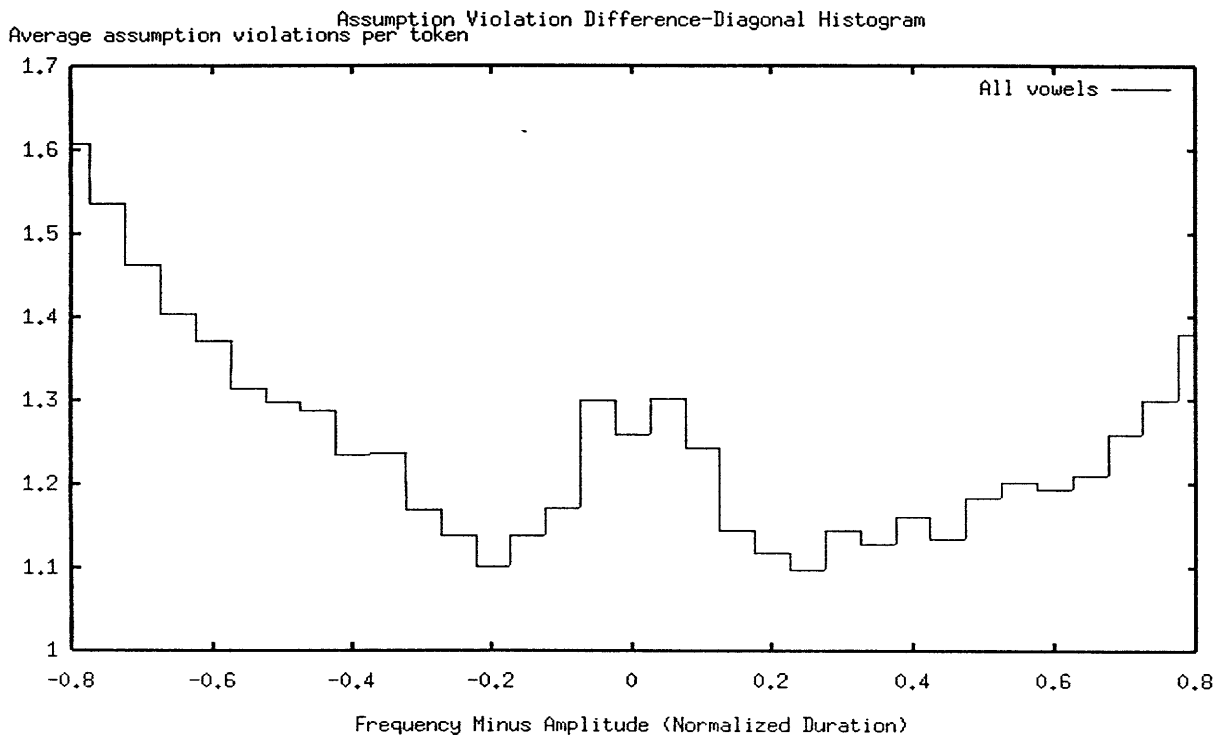


Figure 3-15: Assumption Violation Difference-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the difference between frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the average number of assumption violations per token, for tokens in that bin.

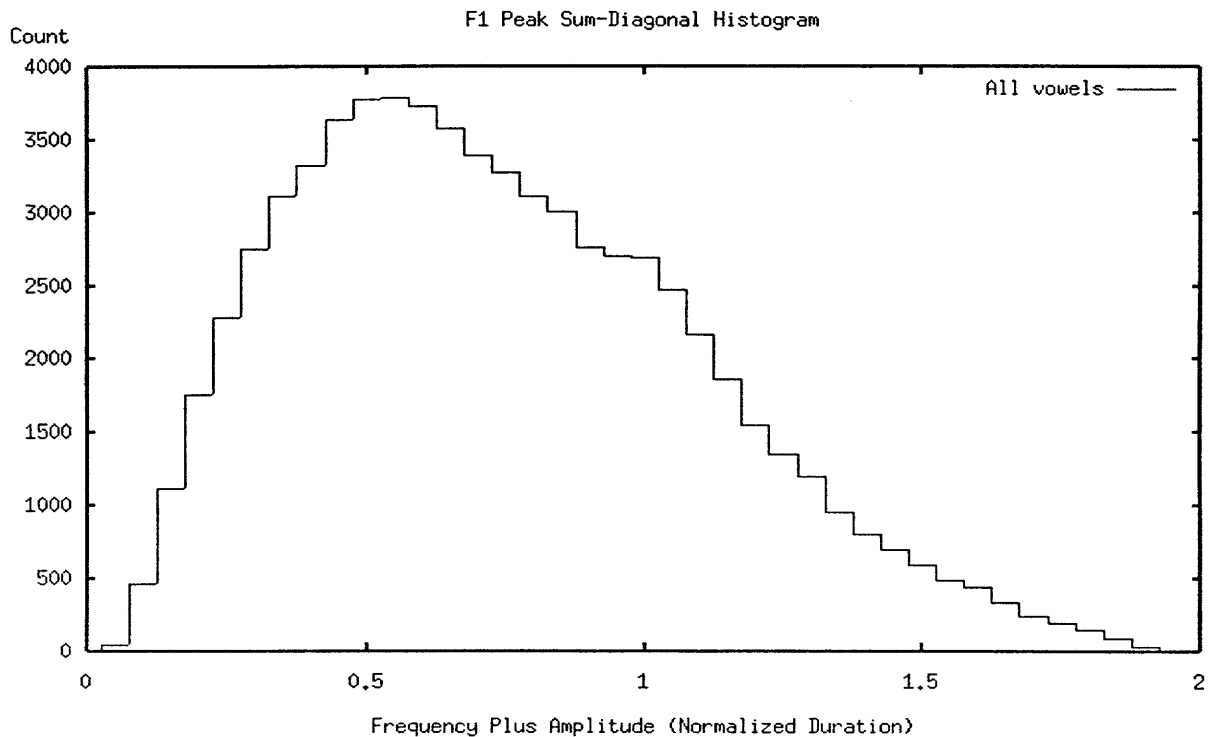


Figure 3-16: F1 Peak Sum-Diagonal Histogram. The data include all vowels in table 3.18. The horizontal axis is the sum of frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin.

is evident from Figure 3-12 that there is a preponderance of vowel tokens at the zero point, but why this should be is not clear.

We can also plot the other diagonal of the cross histogram in Figure 3-10. The result is shown in Figure 3-16. In this case, we see the histogram along the diagonal that represents the average of the frequency and amplitude peaks (absent a scale factor). As in Figures 3-8 and 3-9, this figure clearly shows that peaks tend to occur before the midpoint of the vowel (at least, as the vowel's endpoints are labeled in the aligned phonetic transcription), which is not a violation of any prediction, but is a phenomenon worth noting nonetheless.

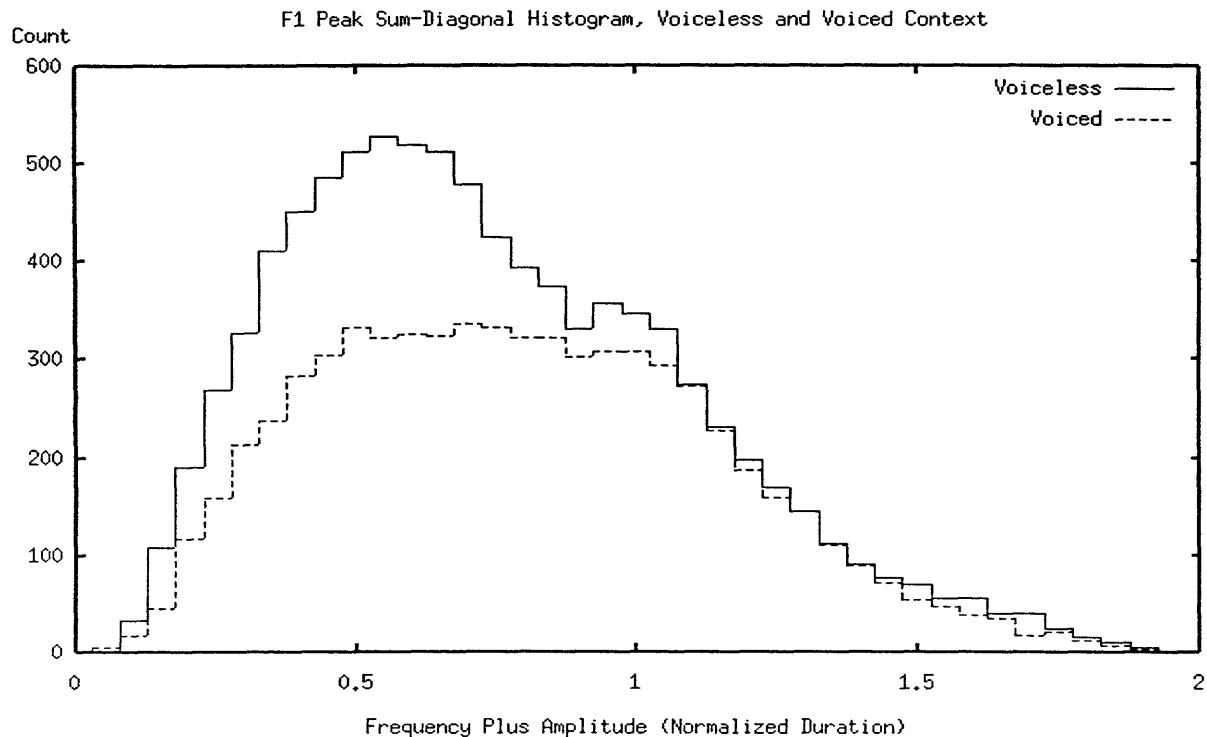


Figure 3-17: F1 Peak Sum-Diagonal Histogram, Voiceless and Voiced Context. The data include all vowels in table 3.18 which are preceded by stop consonants. The horizontal axis is the sum of frequency and amplitude peak locations, normalized against the duration of the vowel. The vertical axis is the number of tokens per bin, computed separately for vowel tokens preceded by voiceless and voiced stop consonants.

One possible explanation of this phenomenon is that, in the TIMIT aligned phonetic transcription, the beginning of the vowel is marked at the onset of voicing energy from the glottal source [47]. This means that a vowel which is preceded by a voiceless aspirated consonant has its beginning marked later than it would be otherwise, presumably later than the beginning of the vowel gesture. The result would be that the center of the vowel gesture would appear before the midpoint of the (labeled) vowel.

To test this idea, the sum diagonal histogram of Figure 3-16 is computed separately for

vowels preceded by voiceless and voiced stop consonants. The result is shown in Figure 3-17. It does appear that voiceless context causes the F1 peak to appear earlier in the vowel than voiced context. However, even in voiced context, the peak tends to appear before the midpoint. Thus, preceding voiceless context seems to account for part of the early skewing of the F1 peak but not all of it. This phenomenon is worthy of further study.

3.7 Experiment 3: Vowel quality better at F1 peak than midpoint

Hypothesis:

F1 peak is a better location for vowel classification than the midpoint of the vowel's duration.

Assumptions:

(a) F1 has in fact a definite peak, not at the endpoints.

(b) F1 is clear and measurable, without interference by phenomena such as nasalization, glottalization, reduction, or aspiration. Exceptions are likely to cause failure of the formant tracking algorithm.

Motivation:

If F1 is pulled down by consonants at the beginning and end of the vowel, it is reasonable to assume that the F1 peak is the place where the vowel is least affected by the surrounding consonants, and therefore it should be a good starting point for vowel recognition.

The results of Experiment 2 show that the F1 peak is usually not at the midpoint of duration of the vowel, and frequently is significantly earlier. This fact is significant because the hand labels for the LAFF database seem to be placed at the vowel's midpoint of duration, in general [10]. If there is a significant difference between a vowel recognizer's performance at the F1 peak and its performance at the midpoint, the LAFF database labeling may need to be revised.

(On the other hand, the main point is to find a landmark somewhere in the vowel. If the vowel classification process is intelligent enough to use the entire extent of the vowel for classification, the precise location of the landmark may not be very important.)

3.7.1 Methodology

The Processed experimental data (see section 3.5.1) were filtered to remove schwas and syllabic sonorants. Out of the total of 80856 vowel tokens, 54060 tokens (about 67%) were retained at this stage.

As in Experiment 2, the data were also filtered to keep only those tokens which have amplitude and frequency peaks both in the middle of the vowel. Out of the 54060 proper vowels, 47576 (about 88%) were retained at this stage.

In addition, all of the shibboleth utterances (sa1 and sa2) were removed from the data, as they have the potential to skew the results of the recognition experiment. (Since all the sa1 and sa2 utterances are of the same sentence, they can skew the statistics of coarticulatory environment.) Out of the 47576 tokens, 37008 (about 77%) were retained.

Finally, the data set was separated into training and test sets, using the division in the

Label	Height	All	Train	Test
iy	high	4925	3546	1379
ih	high	5028	3747	1281
eh	mid	4107	2990	1117
ey	mid	2785	2038	747
ae	low	2967	2227	740
aa	low	2967	2160	807
aw	diphthong	914	704	210
ay	diphthong	2545	1869	676
ah	low	2994	2165	829
ao	mid	2254	1605	649
oy	diphthong	406	286	120
ow	mid	2045	1502	543
uh	high	669	467	202
uw	high	435	329	106
er	mid	1967	1427	540

Table 3.19: Experiment 3 vowel categories and counts.

original TIMIT database. Since the original data were generated by merging the TIMIT training and test sets, this operation restores the original division. Out of the 37008 tokens, 27062 are in the training set, and 9946 are in the test set. Table 3.19 shows the counts of vowel tokens in each category.

Two time points were generated for each vowel token: the midpoint of duration (halfway between the beginning and end points in the TIMIT aligned phonetic transcription), and the peak of F1 amplitude. (The amplitude peak has been shown to be better than the frequency peak for vowel detection, in Experiment 1.)

For each time point, the formant frequency values generated by the formant tracker (see section 3.4.1) form the basic feature set to be used for classification. The formant tracker produces estimates for the first four formants, but only the first three will be used in this experiment.

Since the formant tracker produces a frame of data every 5 milliseconds, the experiment uses the frame nearest in time to the desired time point. In addition, the formant values are linearly interpolated between the 5 millisecond frames, for maximum accuracy in time. Both non-interpolated (time quantized) and interpolated data are used in the experiment, and the results compared.

3.7.2 K Nearest Neighbors (KNN) Classification

Vowel classification is done using a K Nearest Neighbors (KNN) paradigm. This classifier is straightforward to implement and has good performance for large sets of training data [19, section 4.7]. It is rarely used in real-time speech recognition because it is computationally intensive (much more so than many other classifiers), but computation time is not a problem for the experiment at hand. Leung [52] provides a fairly recent example of KNN classification of the TIMIT vowels, which helps inform the present experiment.

To classify a test token using KNN requires a set of N training tokens and a distance metric. The test token is compared to each of the training tokens, using the distance metric, and the K nearest training tokens are found (where K is some number less than N). The token class which appears most frequently amongst the K nearest neighbors is chosen as the output class for the test token.

Besides the training data, this algorithm clearly depends primarily on the distance metric and the choice of value for K . Leung suggests $K = \alpha\sqrt{N}$ based on the theoretical analysis of the algorithm's asymptotic behavior for large training sets [52, pp. 110-116], and reports best results for $\alpha = 1$ [ibid., pp. 118-122].

Leung's feature set (for most of his experiments) is the synchrony spectrum envelope gener-

ated by the auditory model of Seneff [75], which is then normalized for inter-talker differences by shifting the spectrum down (on a bark scale) by an amount dependent on the median value of F0 across the vowel [52, p. 42]. This is intended to reduce the variability of formant locations because of differences in gender and vocal tract size and shape. The result is a 100 element vector of spectral data. Leung uses a Euclidean distance metric for this feature data.

In contrast, the present experiment uses only a 3 element vector (the frequencies of the first three formants). For initial experiments, we will use Euclidean distance as the distance metric. Euclidean distance on the frequency scale would tend to emphasize differences in higher formants, especially F3, since they tend to vary more widely in frequency, so these experiments use Euclidean distance on a bark scale. Initially we will use the value $K = 165 = \sqrt{27062}$.

Extension for K_i

Leung also describes an extension to the basic algorithm, allowing the value of K to vary for different training classes. The K values are computed as before from the size of the training class. $k_i = \sqrt{n_i}$ where n_i is the number of tokens in the i th training class. The decision rule must be normalized by the dispersion of the training class, which (for a Euclidean distance metric) is characterized as a sphere whose radius is the median distance to the k_i nearest neighbors. See [ibid., p. 116] for the details.

First results

Using the non-interpolated (time quantized) data, this KNN vowel classifier yields performance on the order of 50% correct, in most experiments. This is not as good as most of the results reported in the literature over the last ten years, which is not surprising, since those systems are much more complicated than this simple classifier. In fact, the classifier under test performs remarkably well, given its simplicity.

Leung [ibid.] reports about 57% correct for his KNN system, and about 60% correct for his multilayer perceptron, on the TIMIT vowels. Other vowel classification systems yield roughly similar results on the TIMIT vowels.

Meng and Zue [60] used an auditory model to provide a basic spectral representation, optionally followed by extraction of acoustic attributes (spectral moments and amplitude) and estimation of distinctive features, classified by a multilayer perceptron. They demonstrated about 64% correct, slightly dependent on inclusion of acoustic attributes and features. Performance degraded to about 59% when the MLP was replaced by a simpler binary decision.

Chun [13] proposed an hierarchical representation for phonetic classes, with broad decisions made early, followed by finer subclassifications. His system used Mel-cepstrum coefficients and their derivatives as a spectral representation, and Gaussian mixture acoustic models, trained using K means clustering, for classification. On the TIMIT vowels, he reported 67.6% correct for this baseline system [p. 44], but only 52.6% when using the first three formant estimates from the ESPS formant tracker (as my KNN classifier does). Adding F0 to the baseline representation (as an indirect normalization for vocal tract length) improved performance to 70.2% [p. 46].

All these systems report better performance on the TIMIT vowels than the simple KNN

	Midpoint		F1 Peak	
	count	percent	count	percent
Identical	4903	49.3	4513	45.4
High	2136	21.5	2223	22.4
Mid	2304	23.2	2214	22.3
Low	1511	15.2	1424	14.3
Height correct	5951	59.8	5861	58.9

Table 3.20: Experiment 3 vowel recognition rates by height, using non-interpolated formant tracks. Diphthongs are not included in these statistics.

classifier proposed in this experiment. They all use rich spectral information to characterize the acoustic signal, rather than just the first three formants, more sophisticated feature extraction (such as normalization for vocal tract length, principal component analysis, etc.), and powerful classification techniques such as multilayer perceptrons. Our KNN classifier could certainly be enhanced to improve performance, but optimum performance is not the goal of this experiment. The KNN classifier appears to be adequate to characterize the difference between the F1 peak and the midpoint of the vowel.

Results by features [high] and [low]

In addition to identically correct matches, we also examine statistics for how many vowels had the features [high] and [low] recognized correctly. Table 3.19 gives the height values for each vowel label. Table 3.20 shows how many High, Mid, and Low vowels were correctly recognized for their height. Diphthongs are not included in these statistics.

The results in Table 3.20 show that the F1 peak is not as good as the midpoint of the vowel, which disagrees with the hypothesis. However, the difference is slight, and may not be particularly significant. See also Huang's finding that the vowel midpoint is fairly good for classification [38].

	Midpoint		F1 Peak	
	count	percent	count	percent
Identical	4897	49.2	4537	45.6
High	2141	21.5	2213	22.3
Mid	2318	23.3	2243	22.6
Low	1504	15.1	1422	14.3
Height correct	5963	60.0	5878	59.1

Table 3.21: Experiment 3 vowel recognition rates by height, using interpolated formant tracks.

Table 3.21 shows the experimental results when using interpolated formant data. The data are very similar to Table 3.20 in general. As before, performance is usually slightly higher for the midpoint than for the F1 peak, in opposition to the hypothesis.

Why performance should be lower at the F1 peak is a matter of conjecture. It is possible that the formant tracker is making occasional errors, and that some of those errors cause false F1 peak detections which degrade the performance at F1 peaks relative to performance at midpoint (which is not subject to such problems).

From these results, it appears that the midpoint and the F1 peak provide about the same performance for vowel characterization. This is reassuring for the LAFF database labeling project, since it implies that precise location of the Vowel landmarks is not critical.

However, similar performance is not enough to imply that there is no significant difference between the midpoint and the F1 peak for vowel characterization. If the two algorithms are failing on different tokens or in different ways, their difference can be significant even if the failure rates are similar. McNemar's Test [27] is an appropriate test of statistical significance for two algorithms tested on the same data set.

The joint performance of the two algorithms is summarized in Table 3.22. This table shows the number of tokens which both algorithms recognize correctly (upper left corner) or incor-

	Midpoint	
Peak	Correct	Incorrect
Correct	3742	795
Incorrect	1155	4254

Table 3.22: McNemar Test results. The measurements made at the F1 peak are separated on the vertical axis, and the measurements made at the midpoint are separated on the horizontal axis.

rectly (lower right corner), as well as those which are recognized correctly by one algorithm and incorrectly by the other (minor diagonal).⁴ There are quite a few tokens which are recognized correctly by one algorithm and incorrectly by the other (1950 out of 9946, or 19.6%), which indicates that the two algorithms are behaving quite differently.

For McNemar's Test, we compute the probability of this observation under the assumption that the two algorithms are essentially the same. Using the Normal approximation to the Binomial distribution [27, p. 533], the result is very close to zero; that is, there is almost no chance of this observation if the two algorithms are the same. Therefore, the two algorithms are significantly different, even though their performances are similar.

Conclusions

The results of this experiment are rather ambiguous. The significant difference between the F1 peak and the midpoint implies that the location of the Vowel Landmark is important. If so, researchers who are labeling the LAFF database by hand should be aware of the potential impact of Landmark placement. (Apparently, most Vowel Landmarks are being placed at the midpoint of duration, so far.)

However, it cannot be said from these results that one placement yields substantially better

⁴This data shows results for identification only, not by vowel height.

performance than the other, for this experiment. In addition, in any realistic recognition system, the Vowel Landmark will be only a starting point for vowel classification. More detailed analysis of the duration of the vocalic region, prosodic information, and formant movements will be desirable for the detection of diphthongs, vowel - vowel sequences, and semivowels. In this light, the exact placement of the Vowel Landmark ought not to be critical.

3.8 Experiment 4: Fixed energy band is comparable to formant tracking

Hypothesis:

A fixed energy band (without formant tracking) can be found that provides performance comparable to the energy around the frequency of F1 (which requires formant tracking).

Assumptions:

Performance, in this context, means the degree to which peaks appear in the vowels as marked by the TIMIT labeling. It does not mean lack of peaks in the nonvowels, because the formant tracker cannot perform reliably in most nonvowel regions. In other words, we are looking at detection rates, not at insertion error rates. Insertion errors (and their elimination) will be addressed in the next chapter.

Performance can be measured either by the percentage of vowels (as marked in the TIMIT aligned phonetic transcription) which show a proper peak in the energy track, or by the percentage of violations of the theoretical prediction (peak appearing at the end of the vowel

adjacent to an orally closed consonant). Both measures will be used.

Motivation:

Formant tracking is a notoriously difficult task to do reliably, especially under variable conditions (multiple talkers, background noise, and so on). Formant tracking is also computationally burdensome. Since Vowel Landmark Detection is one of the first stages of processing in a LAFF paradigm, we desire a technique which is as simple and robust as possible. The elimination of formant tracking will certainly make the Landmark Detector simpler, and hopefully more robust as well.

3.8.1 Methodology

The experiment seeks to reproduce the results of Experiment 1, looking for peaks in an intensity track derived from a fixed frequency band, rather than the intensity around F1. The intensity track was computed by a weighted sum of spectral bins from the spectrogram. (Of course, the data were converted to linear intensity representation for the summation, and then converted back to decibels.)

A trapezoidal window (in frequency) was used for the weighting operation. It was controlled by four parameters: upper and lower passband edges, and upper and lower transition band widths (all in Hertz, which were converted to bin numbers for the computation). The trapezoidal window and its parameters are shown schematically in Figure 3-18.

For each vowel in each utterance in the TIMIT database, the track was searched for peaks as in Experiment 1, and statistics gathered in the same way.

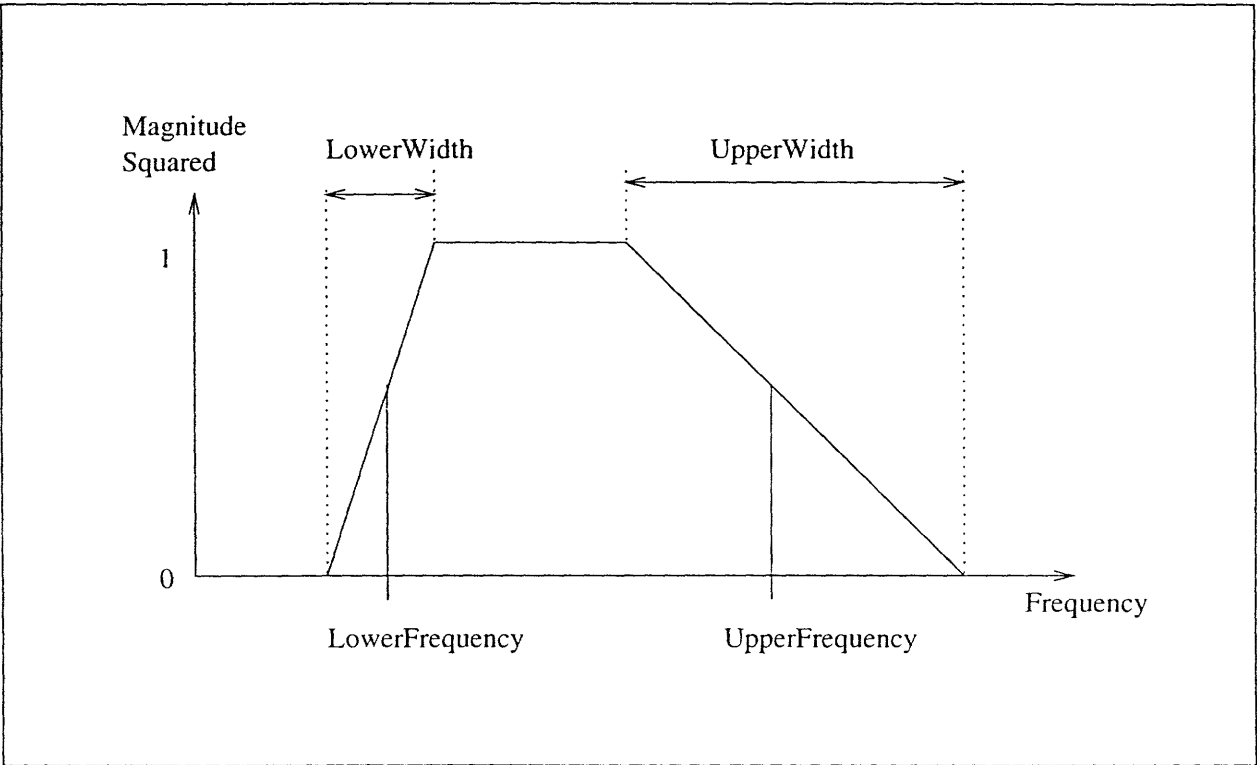


Figure 3-18: Trapezoidal window for weighting.

Unlike Experiment 1, part of the task is to determine what are the optimal parameters for the computation. There are several different values which could be used to quantify performance. For this experiment, two measures of performance were used: percent of all vowel tokens which show a proper intensity peak in the middle of the vowel (which we wish to maximize), and percent of prediction violations (which we want to minimize).

Results

Initially, the values were set as LowerFrequency = 300 Hz, LowerWidth = 0, UpperFrequency = 900 Hz, UpperWidth = 0, which are considered to be fairly canonical for F1 [83]. The statistical results for these canonical values are shown in Table 3.23. Values representing unpredicted results (contrary to theoretical prediction) are printed in emphatic typeface.

The data show performance results which are very similar to the Experiment 1 data shown in Table 3.5. Of the CVC tokens in the database, almost all show an amplitude peak in the middle (27542 out of 27712, or 99.4%). When an amplitude peak appears at the beginning, it is almost always in OV- context (952 out of 978, or 97.3%), and when an amplitude peak appears at the end, it is almost always in -VO context (1711 out of 2115, or 80.9%). These values are comparable to those derived from the Experiment 1 data in section 3.5.2.

Over 96% of all the vowels in the database show an amplitude peak in the middle. This value is even higher than in the Experiment 1 data. The conclusion after this initial test is that the fixed band of frequencies can provide comparable performance without use of a formant tracker.

Optimization

Overall				
	C	O	X	All
C	27712 (34.3%)	9846 (12.2%)	5614 (6.94%)	43172 (53.4%)
O	16020 (19.8%)	5314 (6.57%)	3636 (4.50%)	24970 (30.9%)
X	8361 (10.3%)	2993 (3.70%)	1359 (1.68%)	12713 (15.7%)
All	52093 (64.4%)	18153 (22.5%)	10609 (13.1%)	80855 (100%)
beginning				
	C	O	X	All
C	<i>13 (0.0161%)</i>	<i>0 (0%)</i>	<i>0 (0%)</i>	<i>13 (0.0161%)</i>
O	661 (0.818%)	97 (0.120%)	194 (0.240%)	952 (1.18%)
X	13 (0.0161%)	0 (0%)	0 (0%)	13 (0.0161%)
All	687 (0.850%)	97 (0.120%)	194 (0.240%)	978 (1.21%)
middle				
	C	O	X	All
C	27542 (34.1%)	8859 (10.9%)	5497 (6.80%)	41898 (51.8%)
O	15314 (18.9%)	4788 (5.92%)	3419 (4.23%)	23521 (29.1%)
X	8306 (10.3%)	2698 (3.34%)	1339 (1.66%)	12343 (15.3%)
All	51162 (63.3%)	16345 (20.2%)	10255 (12.7%)	77762 (96.2%)
end				
	C	O	X	All
C	<i>157 (0.194%)</i>	987 (1.22%)	117 (0.145%)	1261 (1.56%)
O	<i>45 (0.0556%)</i>	429 (0.531%)	23 (0.0284%)	497 (0.615%)
X	<i>42 (0.0519%)</i>	295 (0.365%)	20 (0.0247%)	357 (0.442%)
All	<i>244 (0.302%)</i>	1711 (2.12%)	160 (0.198%)	2115 (2.62%)

Table 3.23: Experiment 4 statistical results, for the canonical parameters. For each category, the preceding context is shown on the vertical axis, and the following context is shown on the horizontal axis. Results which violate theoretical predictions are shown in emphatic typeface.

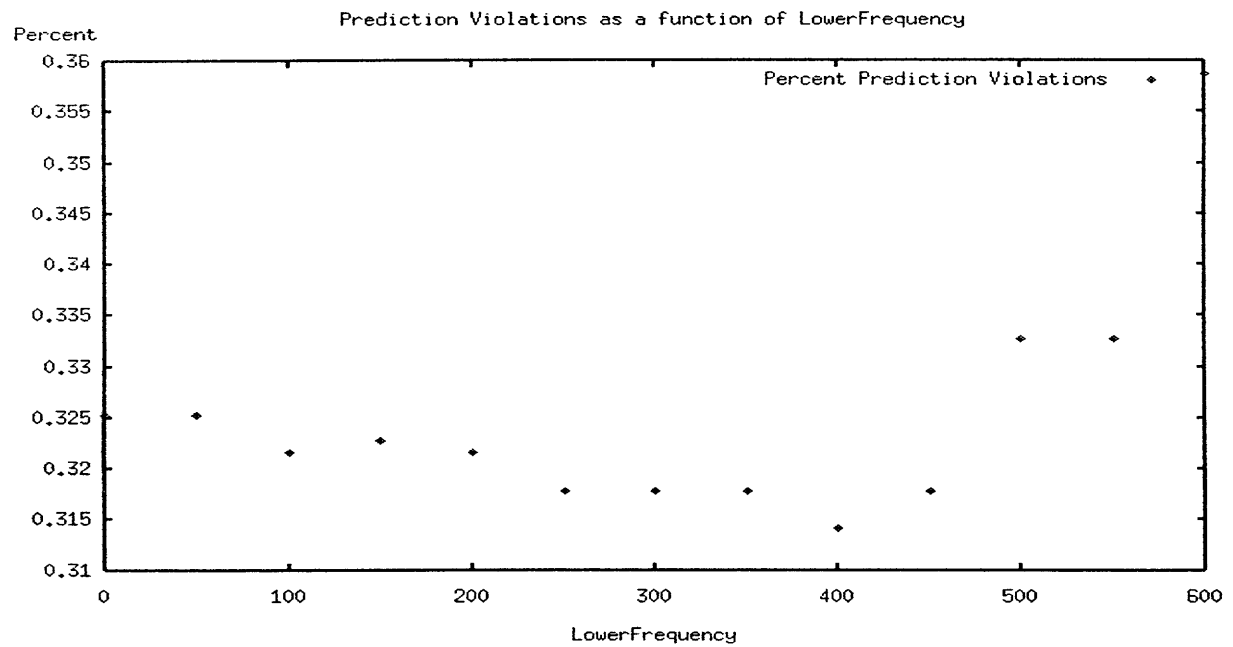
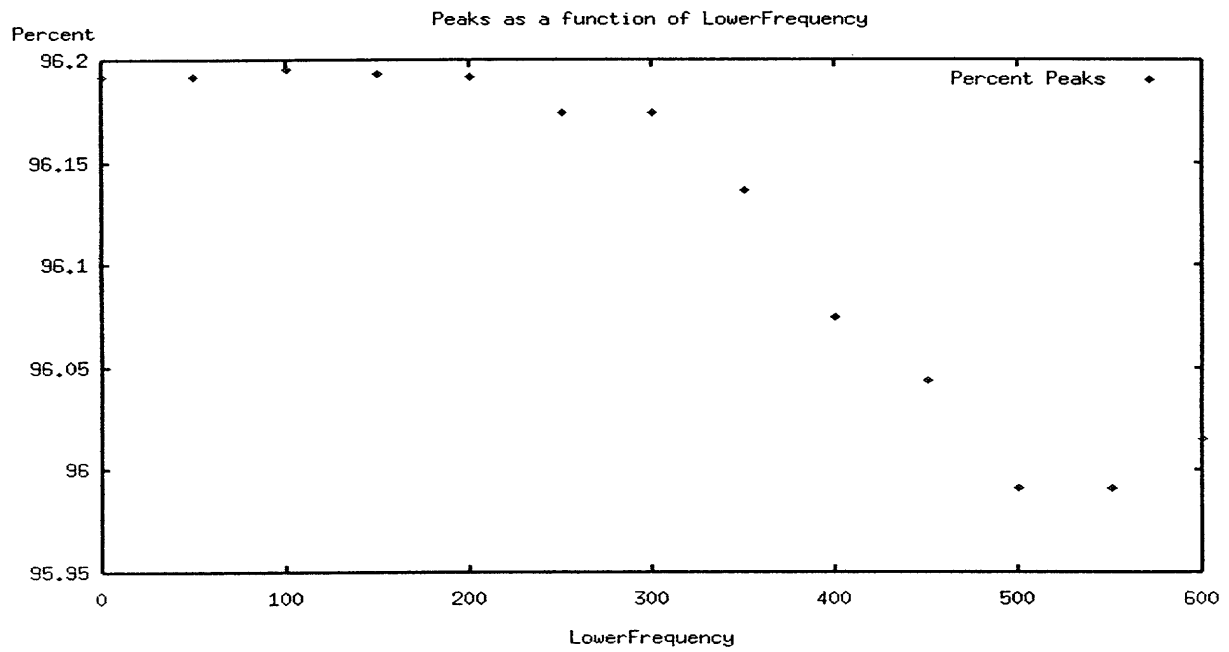


Figure 3-19: Performance as a function of LowerFrequency

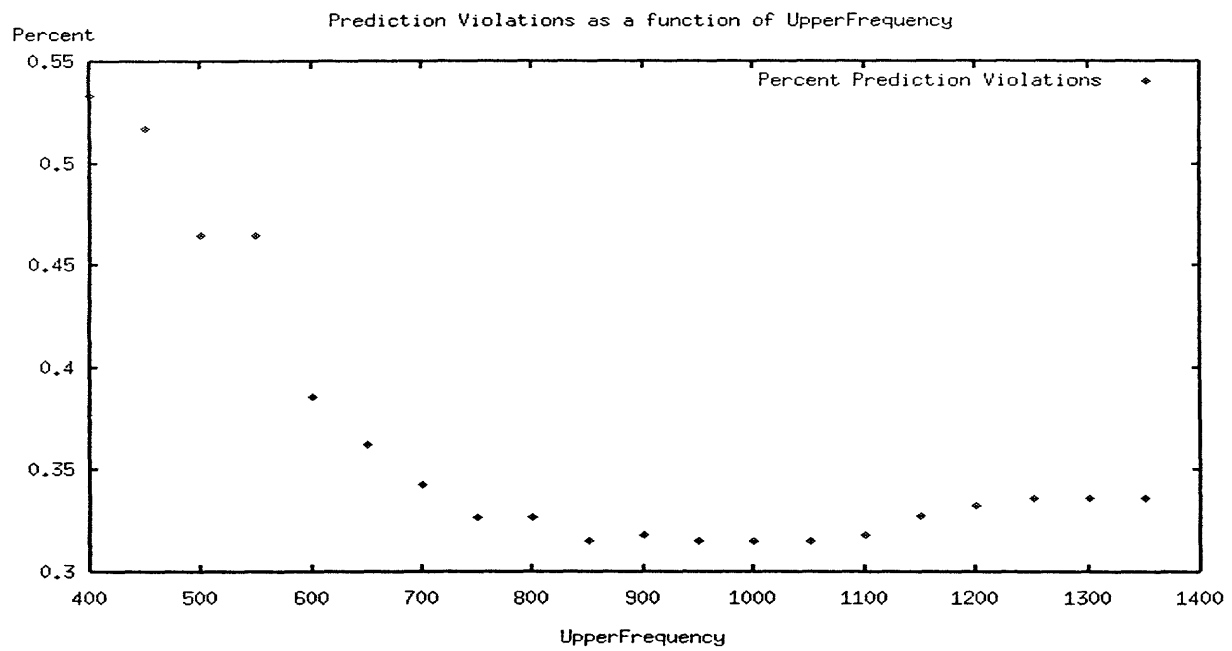
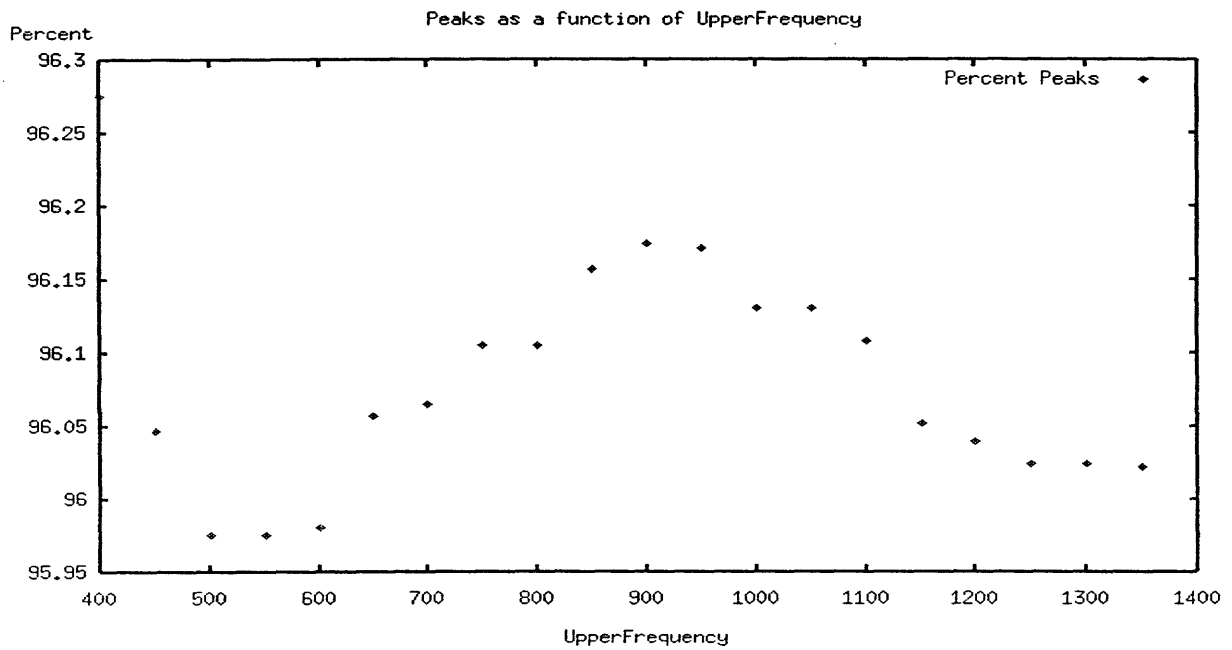


Figure 3-20: Performance as a function of UpperFrequency

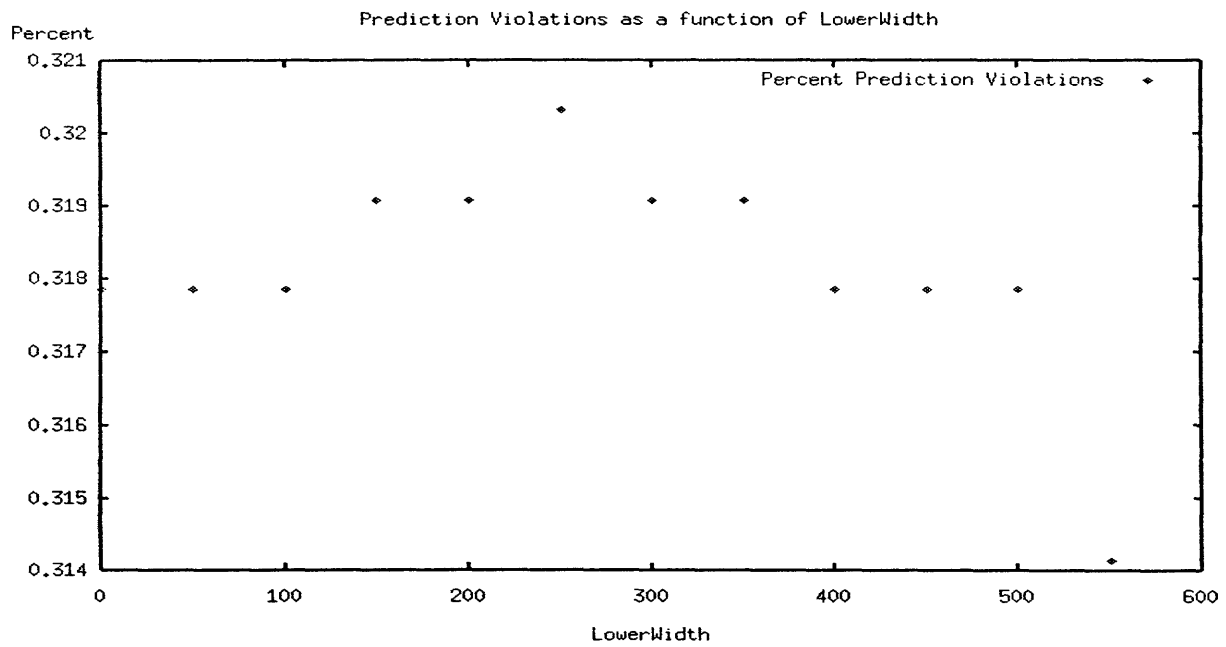
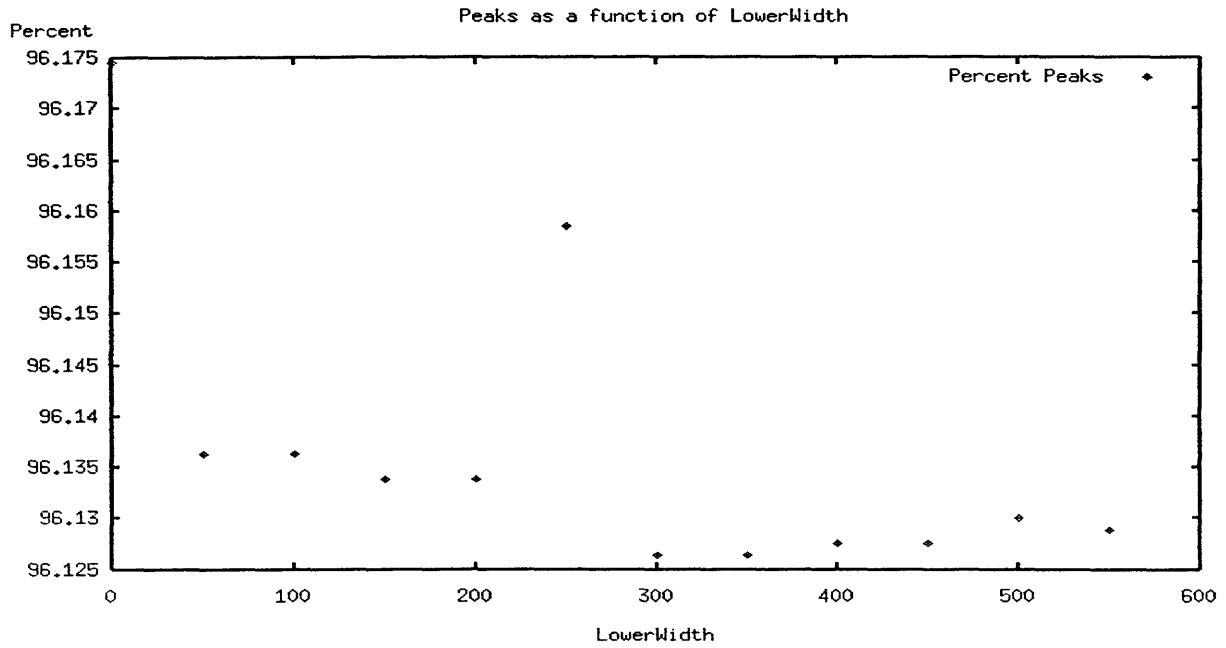


Figure 3-21: Performance as a function of LowerWidth

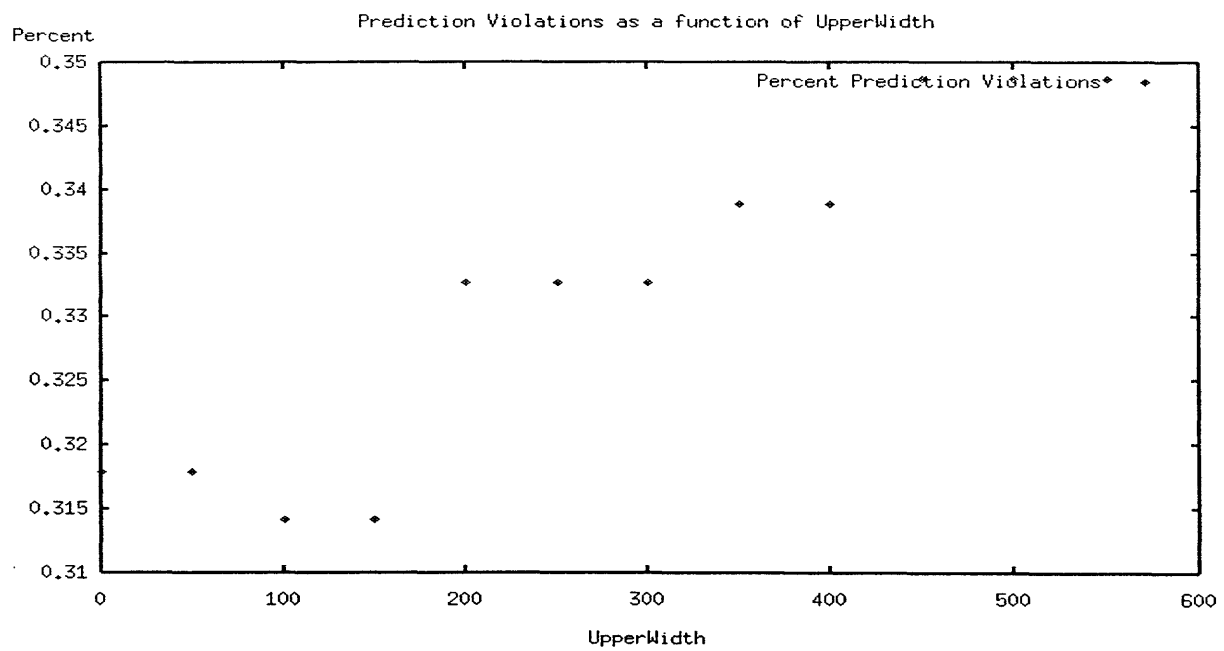
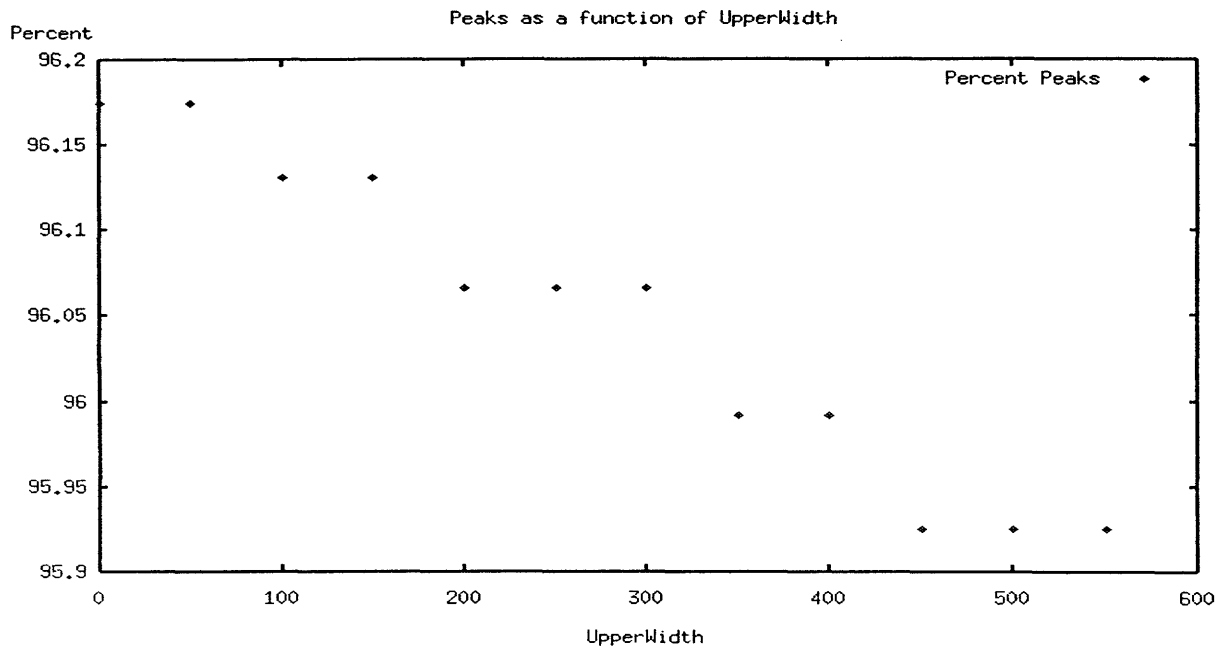


Figure 3-22: Performance as a function of UpperWidth

To investigate further, each parameter value was varied in turn to investigate its effect on performance. Two values were used to gauge performance: the number of all vowel tokens which show a proper intensity peak in the middle of the vowel (which we wish to maximize), and the number of prediction violations (which we want to minimize). Both were expressed as a percentage of the total number of vowel tokens.

Initially, the values were set to their canonical values (LowerFrequency = 300 Hz, LowerWidth = 0, UpperFrequency = 900 Hz, UpperWidth = 0). Each parameter was varied while the other three were kept at canonical values.

LowerFrequency was varied from 0 to 600 Hz, and the results are shown in Figure 3-19 (which shows both proper peaks and prediction violations). UpperFrequency was varied from 400 to 1400 Hz, and the results are shown in Figure 3-20 (which shows both proper peaks and prediction violations).

Both of these figures show performance that is remarkably insensitive to variations in the frequency values. The variations are fairly wide, but performance changes only by fractions of a percent. This is encouraging for a Vowel Landmark detector, as it implies that the precise values for a fixed frequency band are not critical.

LowerFrequency appears to provide best peak detection between 0 and 300 Hz, and fewest prediction violations between 300 and 400 Hz. Therefore, 300 Hz appears to be a good choice for LowerFrequency.

UpperFrequency appears to provide best peak detection between 800 and 1000 Hz, and fewest prediction violations between 800 and 1100 Hz. Therefore, 900 Hz appears to be a good choice for UpperFrequency.

LowerWidth and UpperWidth were varied from 0 to 600 Hz. The results for LowerWidth are shown in Figure 3-21, and they indicate very little sensitivity to the variation, with performance changes only by fractions of a percent. The results for UpperWidth are shown in Figure 3-22, and they indicate slightly more sensitivity (but still very little), with best performance at low values, between 0 and 100 Hz.

3.8.2 Conclusions

The experimental results show that using a fixed frequency band for this task (finding peaks in TIMIT labeled vowels) yields performance comparable to using the formant tracker. The performance is quite insensitive to the exact values for the frequency band edges and transition widths. The results are encouraging because a fixed frequency band is much easier to implement than a formant tracker, allowing simpler and easier implementation of a Vowel Landmark Detector.

3.9 Conclusions of the statistical study

In general, the predictions of acoustic theory are supported by the experiments in this chapter.

3.9.1 (1) Presence of F1 peaks in the vowel

Overall, the experimental results agree very well with the theoretical prediction. As seen in Table 3.5, over 94% of all the vowels in the database show an amplitude peak in the middle.

Amplitude peaks are more consistent than frequency peaks. This supports the practice of using the amplitude peak as the primary indicator of vowel presence.

Of the vowel tokens which do not show an amplitude peak in the middle, all were found to have one or more violations of the assumptions which underlie the theoretical prediction, as seen in Tables 3.7 and 3.8. Most of these conditions were detectable by a simple automatic procedure which examines the token's context and duration, and the rest were detectable by manual inspection.

3.9.2 (2) F1 amplitude and frequency peak together

The experimental results agree with the theoretical prediction. As seen in Figure 3-12, the locations of F1 amplitude and frequency peaks of F1 are strongly correlated. The "shoulder" just below zero on the horizontal axis indicates a tendency for frequency peaks to occur earlier than amplitude peaks.

The exceptional vowel tokens (in which F1 amplitude and frequency do not peak together) tend to violate the assumptions which underlie the theoretical prediction. The greater the difference between amplitude and frequency peaks, the more assumption violations occur on average, as seen in figure 3-15.

On average, the maxima of F1 tend to appear early in the vowel, as seen in Figure 3-16, rather than at the midpoint, for both amplitude and frequency. The effect is more pronounced for frequency peaks than for amplitude peaks. Contextual variations (specifically, the voicing of preceding consonants) account for part of the effect, but not all of it.

3.9.3 (3) F1 peak compared to midpoint for vowel quality

For a simple spectral classification task, the F1 peak does not appear to be substantially better than the midpoint of the vowel. This result holds for both a vowel identity task and a height-only task. Statistical analysis of the classifier indicates that there is a significant difference between the two methods, as seen in Table 3.22. However, the classification performance is not noticeably different, as seen in Tables 3.20 and 3.21.

The implication is that the precise location of the Vowel landmark is not critical. This is good news for the LAFF database labeling project, allowing labelers to be unconcerned with the exact location of Vowel landmarks.

3.9.4 (4) F1 peak can be approximated without formant tracking

The experimental results show that using a fixed frequency band yields performance comparable to using the formant tracker. The performance is not sensitive to the exact values for the frequency band edges and transition widths. The results enable a Vowel Landmark Detector without formant tracking.

Chapter 4

Vowel Landmark Detector (VLD)

Implementation

With the background information about automatic syllable detection, and statistical studies of the TIMIT database, a Vowel Landmark Detector (VLD) was implemented and its performance evaluated. This chapter describes the implementation of the VLD and its performance evaluation.

The approach was to build a baseline VLD first, and then to make a number of modifications to it, in order to gauge the effect that the different modifications have on the detector, their relative utility, and so forth.

As part of the effort, the method used for evaluation of the performance of the VLD was also modified and enhanced to take into account phenomena which can lead to inappropriate error scoring.

Parameter	Value
Lower band edge	300.0 Hz
Lower transition width	0 Hz
Upper band edge	900.0 Hz
Upper transition width	0 Hz
Frames in smoothing average	10 frames
Peak-to-dip threshold	2 dB
Duration threshold	80 ms
Level threshold	25 dB (below overall maximum)

Table 4.1: Parameters for Vowel Landmark Detector, with typical values.

4.1 The Baseline VLD

The baseline detector was designed to reproduce the automatic syllable detector described by Mermelstein [61]. Mermelstein's prototype is a simple knowledge based detector that works by detecting peaks and dips in an intensity track. It appears to be one of the most straightforward and reliable algorithms in the literature, and the most referenced by other researchers.

The VLD is controlled by several parameters which may be adjusted for good performance. The parameters are constant values, and do not change during operation. They include thresholds, frequency band edges, and the like. A list of parameters with typical values appears in Table 4.1. The band frequencies and transition widths are the same as previously shown in figure 3-18.

4.1.1 Front end processing

The first step in processing is to compute a spectrogram of the speech signal. The signal is not preemphasized. The signal is windowed at a frame rate of 200 Hz (or a frame period of 5.0 ms, like most of the algorithms in table 2.1). A 16.0 ms Hamming window is used, with the intent of including at least one full pitch period for male voices. The spectrum is computed from the windowed data with a 256 point FFT. The spectral data are then converted to log magnitude (decibel) representation.

4.1.2 Feature extraction

The second step in processing is to compute an intensity track, summing energy over a band of frequencies to provide a single measure of intensity. Mermelstein's original algorithm used a fairly broad band (500 Hz to 4 kHz, with 12 dB/octave rolloff outside that band). The baseline detector for this thesis, however, was adjusted to track low frequency energy (in the neighborhood of F1, canonically around 300 to 900 Hz). See section 4.3 for a justification of this change. The intensity track is computed by a weighted sum of spectral bins from the spectrogram. (Of course, the data are converted to magnitude squared intensity representation for the summation, and then converted back to decibels.)

A trapezoidal window (in frequency) is used for the weighting operation. It is controlled by four parameters: upper and lower passband edges, and upper and lower transition band widths (all in Hertz, which are converted to bin numbers for the computation).

The resulting intensity track is lowpass filtered in time, to help reduce small peaks and noise. The lowpass filter is a simple rectangular window. One parameter controls the number of frames in the lowpass filter window. In practice, only a few frames (five or so) need to be

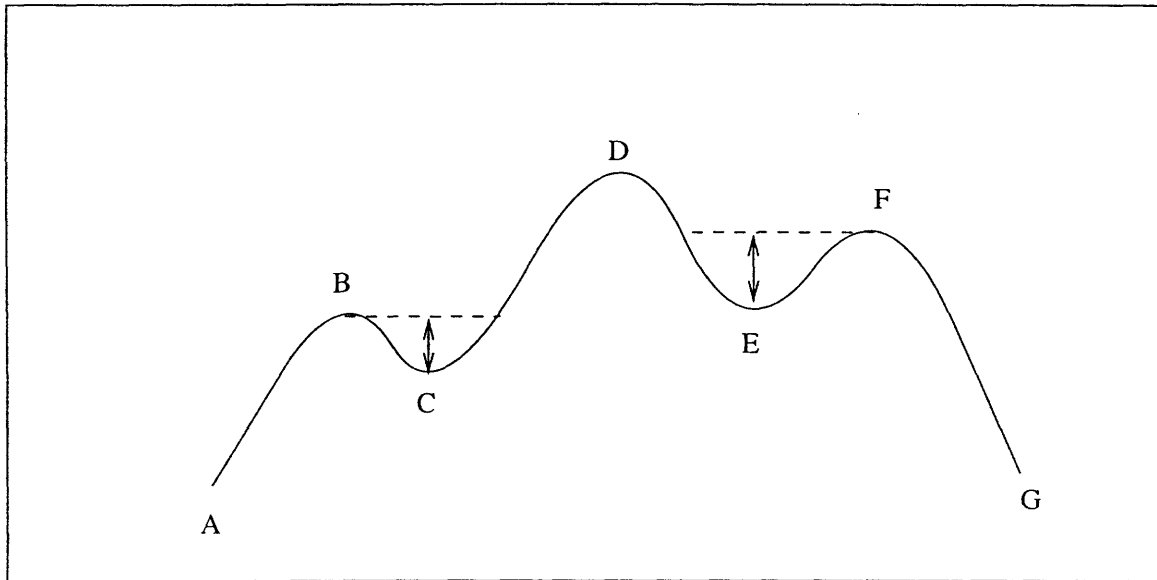


Figure 4-1: Convex hull algorithm (after Mermelstein). See the text for a description of the procedure.

averaged for good results.

4.1.3 Detection

Peaks and dips are detected using a recursive convex hull algorithm [61]. The convex hull is computed by traversing the track from its endpoints inward towards its maximum, maintaining intermediate maxima along the way. The deepest dip is compared to a threshold parameter. If it is deeper than the threshold, the dip is accepted as a boundary, and the process recurses on the two segments thus generated. If not, the recursion ends, and the maximum is accepted as a vowel landmark (or, in Mermelstein's original work, a syllabic center).

For an example, see figure 4-1. The convex hull (dotted line) is computed on the segment A-G. The deepest dip to the left of the peak (B-C) is compared to the deepest dip to the right of the peak (E-F). If the depth of the deeper of the two (say, E) is greater than the threshold, the algorithm divides the segment into two at that dip, and recurses on the segments A-E and E-G. If not, the algorithm returns the peak (D).

Mermelstein's original algorithm also included durational and absolute level constraints.

In order to prevent segmentation into fragments that do not contain adequately strong syllabic peaks, we reject any segment whose intensity maximum is more than a given threshold below the overall intensity maximum, the syllabic peak of the strongest syllable of the utterance. Similarly, a minimum syllabic-unit duration of 80 ms is imposed, and segmentation that would result in shorter fragments is rejected. [ibid, p. 880]

The "given threshold" for level is chosen at 25 dB later in the paper (however, Mermelstein does not give a rationale for either the duration or level cutoff values). These constraints appear to be imposed during the convex hull recursion, as additional conditions which can terminate recursion, but this is not entirely clear from the written description.

None of the later published work using Mermelstein's algorithm seems to implement the duration and level constraints, or indeed to address their existence. In the absence of any other information, the baseline system will implement these constraints as additional conditions which can terminate the convex hull recursion.

4.1.4 Post processing

Mermelstein's original algorithm also included post processing to remove fricative peaks [63]. This was necessitated by the fairly broad frequency band he used (see section 4.1.2), which included the high frequencies characteristic of frication. Our frequency band, encompassing the nominal range of F1, does not include typical frication energy. Therefore, the baseline algorithm did not include post processing to remove fricative regions, except for the very first experiment, which attempted to reproduce Mermelstein's results exactly (as described in section 4.3, which gives the details of the fricative detection process).

4.2 Experimental Issues

The VLD is evaluated by tests which compare its output (the detected landmarks) to an aligned phonetic transcription of the input speech. The relative utility of the extensions and modifications is assessed in the same way. As discussed in section 1.4.3, error rate was chosen to be the measure of performance, with insertion and deletion errors weighted equally. (Other weightings of the errors may be desirable, depending on the needs of the system using the VLD, as was discussed in section 1.4.3.)

4.2.1 Failure modes

There are three general categories of phenomena that could cause the scoring procedure to detect an error (either insertion or deletion).

an effort to avoid scoring artifacts, we introduce the notion of Reference Vowel Landmarks (RVLMs) which are derived from the transcription, and which may be compared to Detected Vowel Landmarks (DVLMs) which are the output of the VLD.

RVLMs are Vowel landmarks, just like the output of the VLD, but they are derived from the database transcription. Different databases have different transcription formats (orthography, aligned phonetic segments, landmarks) which would require different scoring techniques if they were used directly. By converting them to RVLMs, we can use the same scoring technique across all databases, permitting direct comparison of the results. The exact method for deriving RVLMs depends on the format of the database transcription. See section 4.2.3 for details.

The RVLMs are stored in a file which includes Vowel landmarks as well as intervowel dips (roughly corresponding to syllable boundaries). Scoring is done by searching for DVLMs that appear between the intervowel dips for a given RVLM. If none appear, the RVLM is marked as a deletion error, and if too many appear, the extra DVLMs are marked as insertions.

The RVLM format allows for optional matching. Every RVLM must match at least one DVLM to avoid a deletion error. Some RVLMs may be marked to allow matching to more than one DVLM. Such RVLMs include a maximum number N , and may match any number of DVLMs between 1 and N without error. This enables the scoring procedure to avoid penalizing marginal cases, epenthetic vowels, and the like.

At present, we allow optional landmarks only for directly abutting vowels, which may be subject to coalescence (see section 1.3.2). These are the places where the database transcriptions seem to be questionable or inconsistent. Other phenomena, such as vowel deletion or epenthetic insertion, ought to be represented in the acoustic transcription, and we will trust that the transcription is accurate enough for experiments. (If at any point we decide to use



to as the Core Test set in the remainder of this chapter.

4.3 Baseline Experiment

The first experiment started with Mermelstein's original parameters. The object of the experiments was to provide a baseline of performance, against which further changes and enhancements can be compared.

Results

The fundamental value used to characterize performance is Token Error Rate (TER), which is the sum of insertions and deletions as a percentage of RVLMTs (or, in Mermelstein's original work, syllable tokens).

The first experiment attempted to reproduce Mermelstein's algorithm exactly as described in [61]. The algorithm was tested on both the training and test sets (although no training was done). Experimental results are shown in the first row of table 4.2. Overall performance was 26.2% TER, substantially worse than Mermelstein reported (9.5% errors, 6.9% deleted, 2.6% inserted). Part of this difference may be due to the more comprehensive data set. Mermelstein's original work used only eleven sentences spoken carefully by two male talkers, while TIMIT includes hundreds of sentences spoken casually by many talkers of both genders and diverse dialects. There may also be differences in the details of implementation.

Mermelstein's algorithm tracked energy in a frequency range of 500 Hz to 4 kHz, which is marked in table 4.2 as "broadband." It was felt that the broadband intensity measure was

not optimal. The definition of vowel landmarks specifies that they should be located around peaks in energy in the region of F1. If so, the performance should improve when the intensity is measured in a band around F1, nominally about 300 to 900 Hz.

To investigate the effect of this band, the upper and lower band edges and their rolloff values were varied independently. As each parameter was varied, the score (on the Core Training set) was observed, and the best value chosen manually. Then the next parameter was varied in the same way. After several repetitions over the parameter set, convergence to a set of values was observed, and these values were taken as the optimal parameter values.

The optimal frequency band (0 to 650 Hz) does indeed delineate the region where F1 is likely to be found. Performance on this band is 13.4% train, 14.6% test TER (second row of table 4.2). This performance is not noticeably better than the performance using the canonical band (300 to 900 Hz) so either parameter set can be used with essentially the same results.

Mermelstein's original algorithm used post processing for fricative detection, to remove peaks in fricative regions. Mermelstein gives no details of the fricative detection procedure. For this experiment, zero crossing rate was computed and compared to a threshold. When the zero crossing rate around a peak was greater than the threshold, the peak was discarded. Manual adjustment of the threshold, with attention to the score output, resulted in a threshold of 6000 crossings per second, for speech data sampled at 16 kHz.

Observing that fricative regions are much less likely to be detected when using the F1 band, we investigated performance without fricative detection. As expected, performance is very poor when using broadband intensity (third row of table 4.2, 34% TER!), but much better when using F1 intensity (fourth row of table 4.2, 14.8% TER). It appears that fricative detection offers essentially no performance gain when using the F1 frequency band. Therefore,

	Train					Test				
	Tokens	Detect	Insert	Delete	TER	Tokens	Detect	Insert	Delete	TER
broadband	7585	76.2%	2.40%	23.8%	26.2%	4404	77.7%	2.61%	22.3%	24.9%
F1 range	7585	88.4%	1.82%	11.6%	13.4%	4404	87.1%	1.73%	12.9%	14.6%
without post processing for fricative detection										
broadband	7585	87.8%	21.8%	12.2%	34.0%	4404	87.8%	19.7%	12.2%	31.9%
F1 range	7585	88.4%	1.96%	11.6%	13.6%	4404	87.1%	1.88%	12.9%	14.8%

Table 4.2: Scores by frequency range, with and without fricative detection. The “broadband” condition is Mermelstein’s original frequency range (500 Hz - 4 kHz), and the “F1” range is 0 - 650 Hz.

	Train					Test				
	Tokens	Detect	Insert	Delete	TER	Tokens	Detect	Insert	Delete	TER
Overall	7585	88.4%	1.96%	11.6%	13.6%	4404	87.1%	1.88%	12.9%	14.8%
Tense	3168	91.7%	3.22%	8.27%	11.5%	1838	89.6%	2.99%	10.4%	13.4%
Lax	1704	92.8%	0.59%	7.22%	7.81%	1001	91.7%	1.10%	8.29%	9.39%
Schwa	2458	81.9%	1.42%	18.1%	19.5%	1446	81.2%	1.11%	18.8%	19.9%
Sonor	255	80.4%	0.784%	19.6%	20.4%	119	83.2%	0.00%	16.8%	16.8%

Table 4.3: Scores by vowel stress. In general, less stressed vowels are more difficult to detect. The exception is lax vowels, which are easier to detect in context (because they are always followed by consonants).

all subsequent experiments were done using the F1 band without fricative detection.

Details of the detector’s performance by vowel stress are shown in table 4.3. As one might expect, full vowels (tense and lax) are easier to detect than schwas or syllabic sonorants. Lax vowels are the most easily detected, even more so than tense vowels, which may result from the fact that lax vowels must be followed by consonants, whereas tense vowels may not be. The consonants provide clear boundaries for segmentation, making detection easier.

In all cases, performance on the test set is very close to performance on the training set, which indicates that this amount of data is more than adequate.

4.4 Combination of Acoustic Measurements

The decision to terminate the convex hull recursion is based on three values: the peak-to-dip level difference, the duration of the segment which would be created, and the absolute level of the peak (relative to the overall peak level). Mermelstein combines these three values in a very simple way – each is compared to a threshold value, and if any one of the three does not meet its threshold criterion, the recursion terminates. This is what was done in the baseline experiment, as described in section 4.1.3.

However, this simplistic combination may not adequately represent the interdependence between these values. For example, a short duration is a fairly good indication that the segment is not a vowel, but a long duration is not a good indication that the segment is a vowel (as in pauses, prepausally lengthened murmurs, and so on). In this case, we want a short duration to terminate the recursion, but a long duration to cause more weight to be given to the other values.

On the other hand, a high absolute level is a good indication that the segment is a vowel, but a low level is not a good indication that the segment is not a vowel. Reduction and devoicing can cause levels low enough to be comparable to nonvowel phenomena such as murmurs or velar stop bursts. The evidence in the absolute level measurement seems different in polarity from the evidence in the duration measurement, undermining confidence in Mermelstein's Boolean test. Furthermore, there may be subtler interactions between the values which are not immediately obvious to intuition.

This issue becomes even more important if we recall that we want the VLD to generate some kind of quantitative measure of strength or confidence in the Vowel landmarks. Mermelstein's scheme implements only a binary decision, yes or no. It is not at all obvious how to combine the three values into a single quantitative score, and even less obvious how to demonstrate

that the computation is optimal in any meaningful sense.

4.4.1 Using intuitive nonlinear combinations

Some time and effort was spent on attempts to compute nonlinear combinations of the acoustic measurements, and then to optimize the coefficients of the combination for best results. The general approach was to normalize each measurement using a linear transformation ($y = ax + b$) and then to pass each normalized measurement through a saturating nonlinearity (such as the arctangent function) and sum the results. The hope was that the transformation coefficients could be optimized using manual inspection or automatic methods. Unfortunately such optimization turned out to be very difficult, primarily due to a lack of a principled process to improve performance from a given starting point. For example, there seems to be no way to demonstrate that the basic architecture (linear transformation of each measurement, saturating nonlinearity, and addition) either is or is not the best method to use for this combination. For instance, when the duration of a candidate vowel is very short, the absolute level may be more important (to distinguish schwas from sonorant consonant phenomena), but it does not follow that when duration is very long, absolute level is unimportant. The conclusion was that there was insufficient insight into the interrelations between parameters to develop a justifiable strategy.

4.4.2 Using Neural Networks

In order to explore the possible nonlinear dependencies between the values in a more principled way, a multi layer perceptron (MLP), one kind of neural network [2], was used. MLPs are capable of discovering nonlinear relationships among their inputs and making optimal decisions based on training data.

The MLP results will serve as an existence proof for the performance we can expect, detecting Vowel landmarks with these three parameters. It is possible that insight may be gained by examining the weights of the trained network, which can then be used to guide the design of an explicit decision stage. The performance of the rule-based decision stage can then be compared to the MLP performance, to see how close it can get. It is also possible to build the MLP into the Vowel landmark detector, and use it to make the decision directly.

There is some philosophical difference of opinion about the use of “ignorance models” (statistically based techniques) in speech recognition. This author believes that ignorance models are entirely appropriate for use when we are ignorant about how information should be used. The widespread use of wholesale ignorance modeling in commercial speech recognition neglects the substantial knowledge that we have about speech through linguistic study and acoustic theory. Indeed, a central motivation for the LAFF paradigm is to incorporate linguistic knowledge and acoustic phonetics into a viable speech recognition technique. However, such knowledge is still incomplete, and ignorance modeling is a useful and appropriate method for filling in the gaps in our knowledge.

Methodology

Convenient tools are readily available for applying neural networks to database detection and classification problems [85]. These tools require a database of input values and output targets, for training and testing. In order to create a database suitable for use with these tools, the VLD of section 4.3 was used in a slightly modified form.

The detection parameters were modified from their optimal values, in the direction of leniency, so that the VLD overgenerated landmarks to a large degree. The parameters chosen were $\text{PeakToDip} = 0.5$ dB, $\text{Duration} = 20$ ms, and $\text{Level} = 50$ dB below peak. The purpose

	Train	Test
Files	619	373
Vowels	8104	4709
Landmarks	9953	5830
Detections	6471	3760
Deletions	1633	949
Insertions	3482	2070

Table 4.4: Basic statistics for the Vowel Landmark Detector, using very lenient parameters to minimize deletion errors.

was to minimize deletion errors, even at the expense of a great increase in insertion errors.

The modified VLD was run on the Core training and Core test subsets of the TIMIT database (described in Section 4.2.3). Basic statistics on the Vowel landmarks generated by this procedure are shown in Table 4.4. It is evident from the table that, even though the parameters were adjusted to minimize deletions, there are still a fair number of deletion errors (20.1% in both training and test sets). We may assume that most of these deletion errors will remain, regardless of the choice of parameters. Although the results of the experiment will not be entirely definitive, we will proceed on the assumption that most of these deletion errors are unavoidable, and may be regarded as constant against variations in the detection parameters.

For each landmark, four values were stored in the database: the time index of the landmark (seconds), the peak-to-dip difference (dB), the segment duration (milliseconds), and the absolute level of the peak (dB below the strongest peak in the sentence).

To generate the targets, the VLD's output landmarks (DVLMs) were compared to the TIMIT database labeling. Each landmark that fell within a labeled vowel was given a target value of +1.0, and each landmark that did not was given a target value of -1.0. The goal was to train a network which would generate high values for vowels and low values for nonvowels.

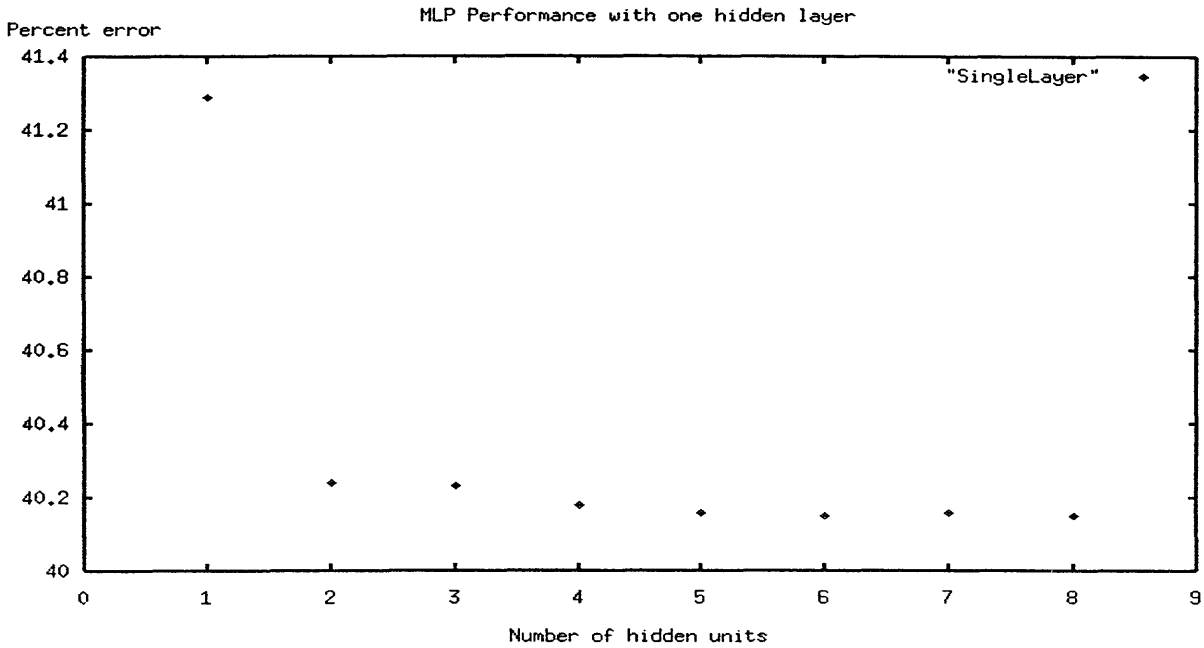


Figure 4-2: Experiment 5 MLP Performance by number of hidden units

A neural network was created, having four input units and one output unit. Because the task required only a static combination of the inputs (without dynamics or state information), a multi layer perceptron (MLP) architecture was chosen. MLPs have only feed-forward paths, without feedback loops or state variables.

MLPs may have one or more hidden layers of units, in addition to the input and output units. One hidden layer is enough to capture very general nonlinear relationships between the inputs, although several hidden layers may be more efficient. Since this task is a very simple one as MLP applications go, one hidden layer was used.

Results

The MLP network was trained on the Core training set (described in section 4.2.3). With one hidden layer, the primary design variable was the number of hidden units. Training was done on versions of the MLP with different numbers of hidden units (one through eight), and the results are shown in Figure 4-2. This figure indicates that one hidden unit is not enough to capture the complexity of the data, and that more than two hidden units do not provide substantial benefit. Therefore, two hidden units were used for subsequent experiments.

Also, the time index was removed as an input. The time index provides information about where in the sentence the vowel appears (since each utterance in the TIMIT database is one complete sentence). An actual VLD in “field” conditions will not have sentence boundary information to work with, and therefore no time index.

The MLP without time information was retrained on the Core training set. Without the time information, error rate increased slightly (from 40.2% to 40.6%), which may reflect the slight decrease in average level over the course of the sentence (see Figure 3-3).

The network values after training are shown in Figure 4-3. The upper part of the diagram shows the input parameter normalization, whose function is to transform the input parameters so that they have uniform distributions (in this case, zero mean and unity standard deviation). The values are based on the mean and standard deviation of the data in the training set. For example, the duration values in the training set have a standard deviation of 63.5 ms, and when divided by this value, the result has mean of 0.56 (or 35.6 ms). In this case, the duration is the length in milliseconds of the subsegment when recursion is terminated, i. e. the first subsegment which does NOT appear to be a vowel. (This is not the vowel duration.)

These statistics implicitly assume a Gaussian distribution, which may be a good approximation for the Duration parameter, since it does not take values close to zero. A Gaussian can

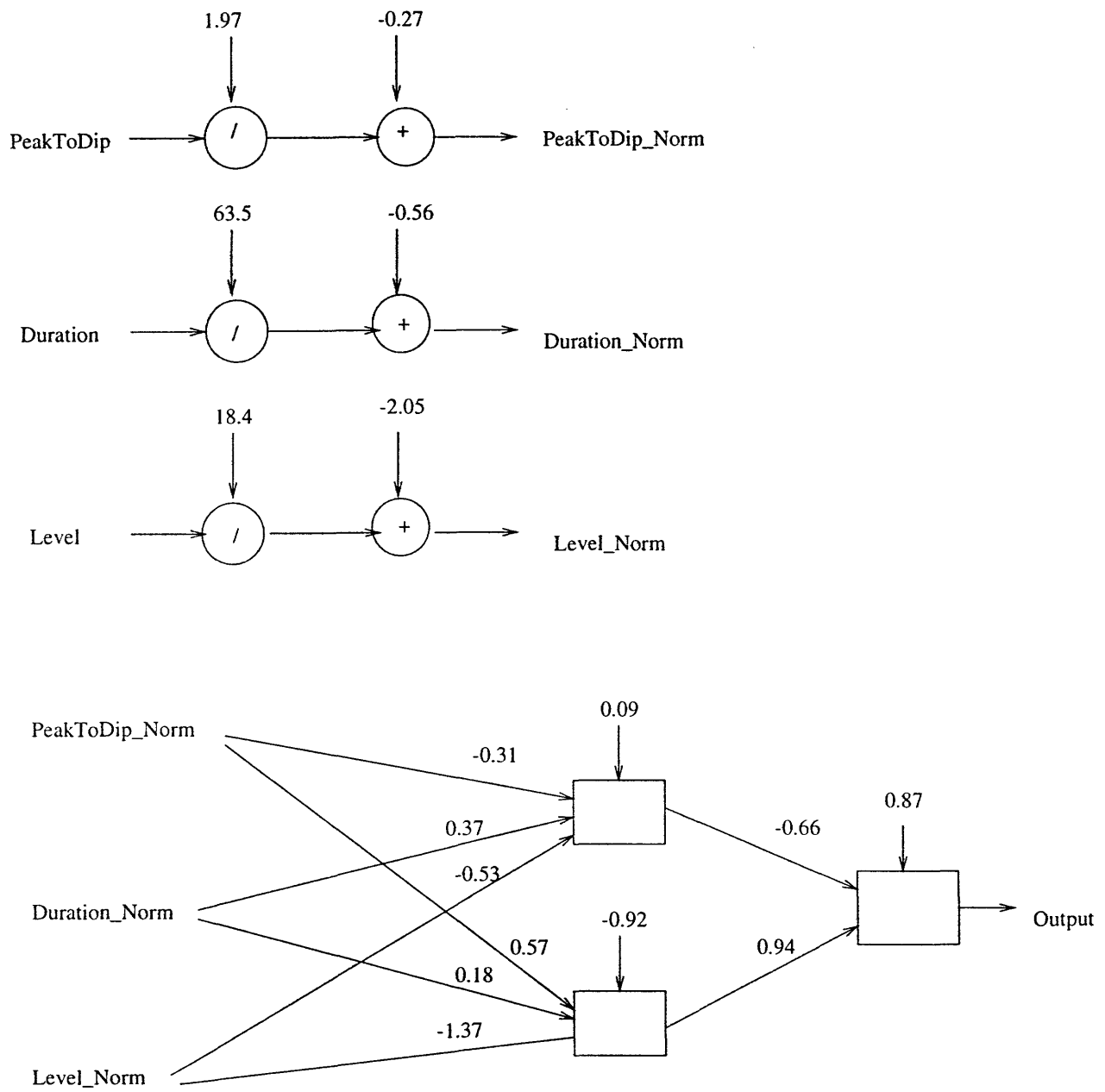


Figure 4-3: Experiment 5 MLP network weights. These weights result from training using back propagation on a sum-of-squares error criterion. See the text for interpretation of the values.

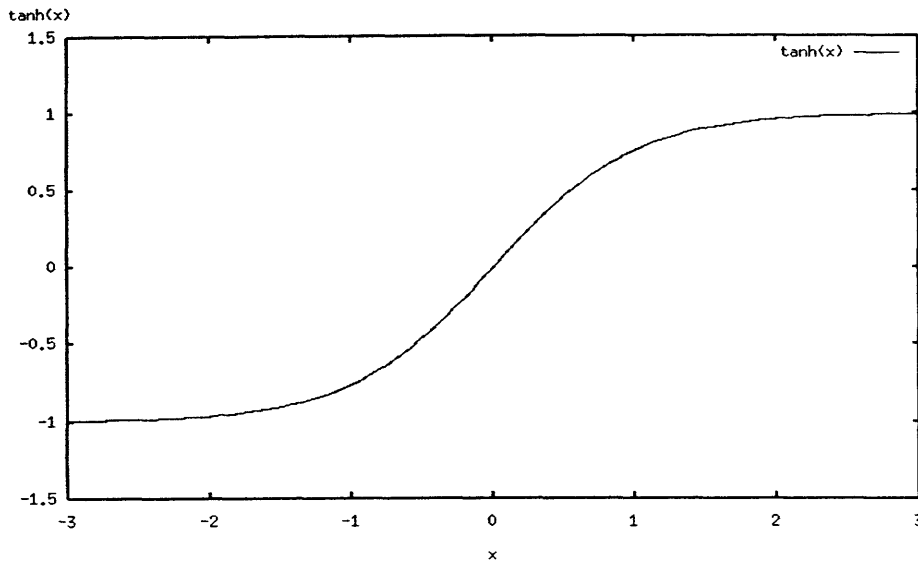


Figure 4-4: Hyperbolic tangent, which is used as saturating nonlinearity for MLP units.

only approximate the distributions of the PeakToDip and Level parameters, which take non-negative values down to and including zero. We will assume the approximation is adequate for now.

The lower part of the diagram is the network proper. Each unit is represented by a rectangle. The unit's function is to sum the unit's inputs (weighted by the connection weights as shown) and pass the result through a saturating nonlinearity. In this case, the nonlinearity is the hyperbolic tangent, which has unity slope at the origin, and saturates to within 5% of unity at around +/- 2.0 (see Figure 4-4).

The output is strongly activated by a positive output from the lower hidden unit (0.94) and less strongly activated by a negative output from the upper hidden unit (0.66). We will call the lower hidden unit the "Yes" unit (evidence that there is a landmark), and the upper hidden unit the "No" unit (evidence that there is no landmark). The bias means that the

output tends to be positive unless strongly negatively activated (0.86).

Recall that Level is the level of the vowel peak in dB below the overall highest level, so that high Level values indicate low level peaks. The “Yes” unit is strongly activated by a high absolute level (1.37), less strongly activated by a high PeakToDip value (0.57), and not strongly activated by a long duration (0.18), with a bias of about -1.0 meaning that there must be at least +1.0 of evidence before this unit achieves positive activation.

The “No” unit is somewhat activated by a high absolute level (0.53), somewhat less strongly activated by a long duration (0.37), and by a low PeakToDip value (0.31), with essentially no bias.

Interpreting these values, even in such a small network, is a rather vague and uncertain task. It seems surprising, for instance, to see high absolute levels activating the “No” unit, since we expect high level to indicate the presence of a vowel, not its absence. Presumably the negative effect of level via the “No” unit is overwhelmed by the positive effect of level via the “Yes” unit. It is important to keep in mind that designating the units “No” and “Yes” are merely approximations to their true functions in the network, and the contribution of each input parameter is spread throughout the network.

However, we can draw some tentative general conclusions. A high absolute level and a high PeakToDip value are strong indicators of the presence of a Vowel landmark. A short duration is a good indicator of the absence of a Vowel landmark, but a long duration is not a good indicator of the presence of a Vowel landmark.

4.5 Incorporating Neural Network into Vowel Landmark Detector

In the previous section, we showed that a neural network, specifically a multi layer perceptron (MLP), can be used to classify landmarks which are output from the VLD, and that only a small network is necessary to do this (two hidden units). The performance values are rather poor in that section, because the sum-of-squares error function does not capture quite the same information as error rate.

4.5.1 Issues

As the MLP is a general and powerful tool for capturing nonlinear dependencies, and making decisions thereby, it is reasonable to consider incorporating this network directly into the VLD as its final decision making apparatus. The optimum values of the MLP can be incorporated directly into the VLD, but even so, we will want to be able to train the network *in situ*, as an integral part of the VLD.

The experiment demonstrates the effectiveness of training techniques; however, there are two issues that must be addressed in order to integrate the MLP into the VLD. Both issues have to do with the choice of error function to minimize.

Back propagation and choice of error function

In order to optimize the MLP, we must have an error function to minimize. The experiment has used the sum of squares of the differences between the target and the MLP output, which

is a fairly common and traditional error metric. However, it does not seem to be completely appropriate for this task.

With the target values set at +1 for “vowel” and -1 for “not vowel,” the sum of squares error metric will weight output values close to zero as substantially worse errors than output values close to unity (of the appropriate polarity). However, this does not entirely capture the desired behavior of the VLD. If the desired result is correct detection of vowel landmarks, an output value of +0.1 is just as good as a value of +0.9 (assuming the decision threshold is at zero).

Error rate (insertions + deletions) would seem to be a better choice of error function than sum-of-squares, for this task. One problem with error rate is that it is not differentiable with respect to the connection weights. (As the connection weights are varied, individual target landmarks in the test change between “detected” and “not detected,” causing discontinuities in the error rate, which is therefore not differentiable.) The back propagation algorithm requires the error function to be differentiable with respect to the connection weights, and so back propagation cannot be used to guide the optimization of the network.

Fortunately there is an alternative. Numerical differentiation can be used to guide the optimization of the network. This is a straightforward technique in which each weight is perturbed slightly in turn, and the resulting changes in the error function are combined to provide a gradient to guide optimization. The most straightforward way to optimize the network is to take a small step in the direction of the gradient and repeat the process (a method we will refer to as gradient descent).

Numerical differentiation (in particular, gradient descent) is generally eschewed in favor of back propagation because it is more computationally burdensome ($O(n^2)$ rather than $O(n)$, where n is the number of connection weights). Computational load is not a concern in this

case because the network is so small (eleven connections, versus the thousands that are usually seen in speech recognition tasks).

If circumstances dictate, it is reasonable to begin optimization using back propagation and sum-of-squares to get close to the desired optimum, and then finish the job using numerical differentiation and error rate to find the best value.

Nonlinear optimization and choice of error function

There is another problem with error rate as an optimizable error function. Almost all nonlinear optimization techniques assume a function which is smooth on a fine scale, in order to find the appropriate gradient for descent. Error rate is not smooth on a fine scale (as observed above, small discontinuities in the error rate occur as individual target landmarks change state with changes in the detection parameters).

It is the author's experience that the Error rate function tends to appear smooth on a large scale, with the discontinuities limited to a small "fuzz" of variation on a small scale. In this case, gradient descent (or similar nonlinear algorithms) may be expected to converge to the general neighborhood of the optimum, but not to converge on the precise location of the optimal point.

In this case, we can use simulated annealing to avoid convergence on a false optimum. This procedure adds a small amount of jitter at intervals to the gradient descent, in order to avoid convergence to a local extremum. Since it is also likely that gradient descent will not completely converge, due to fine scale irregularities, the procedure may be finished by a manual adjustment for best performance.

4.5.2 Implementation and validation

The task was to integrate the MLP into the VLD. In order to do this, a simplified MLP algorithm was coded in C (since the full functionality of the NICO toolkit [85] was not necessary). To validate the code, the weights that resulted from the training on the Core Training set (as shown in Figure 4-3) were installed in this network, and this network was run on the Landmark data (as shown in Table 4.4). This network's outputs matched the NICO network's outputs to within the limits of numerical representation, indicating that the implementation is valid.

The VLD's hard limits were kept in place, so that the MLP only processes segments which pass the hard limits (the limits were kept at their very lenient values, just as in section 4.4.2). Therefore, the MLP operates on the same information as in section 4.4.2, so that the weights from that experiment can be transferred to this implementation directly. These weights (which were trained using back propagation to minimize sum-of-squares error) are not the ideal weights for the final application (which seeks to minimize error rate), but it was hoped that these weights would serve as a starting point reasonably close to optimal.

4.5.3 Training algorithm

The weights of the MLP were trained using simple gradient descent on the error rate measure. Considering the starting set of weights as a point in N-dimensional space (here $N=11$), a fixed increment was used to examine nearby points (plus and minus in every dimension, for a total of $2N$ points). The error function was Token Error Rate (TER), as described in section 4.3. The point with the best error rate was chosen as the new center point, and the process was repeated.

When the best error rate was found to be at the center point, the increment was reduced by half, and the procedure continued. In the first few rounds of training, it was found that once the increment was lower than about 0.005, the center point was always found to be best (apparently, this number corresponds to the smallest feature size of irregularities in the error function). Therefore, the procedure was terminated when the increment was reduced to below $1.0e-3$ and the resulting center point was kept as the final point.

Annealing was done by adding a random value to each weight. The random numbers were generated by the Perl *rand()* function (seeded with time of day), and scaled to the range $[+R, -R]$ where R was a fixed value.

4.5.4 Training results

The weight set generated in section 4.4.2 was used as the first point. The starting increment and the annealing range were both set to 0.1 for this experiment. After training, the resulting point was annealed and retrained eight times. Results are shown in Table 4.5.

Since each point is a vector of N dimensions (here $N=11$), comparisons between the points are difficult to represent spatially. Table 4.5 shows the Token Error Rate (TER) for each point, as defined in section 4.3, and the distances between points in N-space. It is evident that the original weights (which were trained to minimize sum-of-squares error) do not minimize the error rate, as all the training runs result in substantially lower error rates than the original (by about a factor of three).

The distances between early points and later points increase, in almost all cases, with each annealing run. We can interpret this as the descent of a rather bumpy valley in the error function, where each training run converges on a local minimum, and the annealing serves

Point	TER	Distances								
		(%)	Orig	T1	A1	A2	A3	A4	A5	A6
Original	36.1									
Train 1	12.8	1.54								
Anneal 1	12.6	1.60	0.186							
Anneal 2	12.5	1.76	0.501	0.329						
Anneal 3	12.6	1.75	0.618	0.473	0.220					
Anneal 4	12.7	1.86	0.709	0.557	0.286	0.209				
Anneal 5	12.1	1.92	0.945	0.831	0.616	0.463	0.409			
Anneal 6	12.2	2.03	1.00	0.906	0.704	0.572	0.505	0.233		
Anneal 7	11.9	2.00	1.10	1.00	0.814	0.698	0.601	0.437	0.422	
Anneal 8	12.0	2.07	1.19	1.09	0.912	0.811	0.698	0.557	0.504	0.167

Table 4.5: Training results for the VLD using MLP for decisions, for the first eight annealing runs. Token Error Rate (TER) is as defined in section 4.3. Distances are Euclidean distances in coordinate space.

to push the point out of the local minimum so it can continue down the valley. Although there are many local minima, the performance does not vary strongly between them, i. e. the valley bottom is fairly flat, and therefore rather insensitive to the exact choice of weight values.

Since it appears that the valley bottom was not reached in the eight annealing runs of Table 4.5, the series was continued, starting with the last point in Table 4.5. The starting increment and the annealing range were both kept at 0.1 for this experiment. The current point was annealed and retrained a total of sixty-four times. The current point's distance from the original starting point (for all sixty-four runs) in Figure 4-5.

It appears from this plot that the training algorithm takes between 20 and 30 runs to converge on an optimal region, and stays in that region thereafter, but with enough variation that this seems to be a fairly stable optimum. After the first few runs, the error rate stays right around 12 percent Token Error Rate (TER), plus or minus a tenth or two.

The ten points with the best TER performance in Figure 4-5 were selected, and their inter-

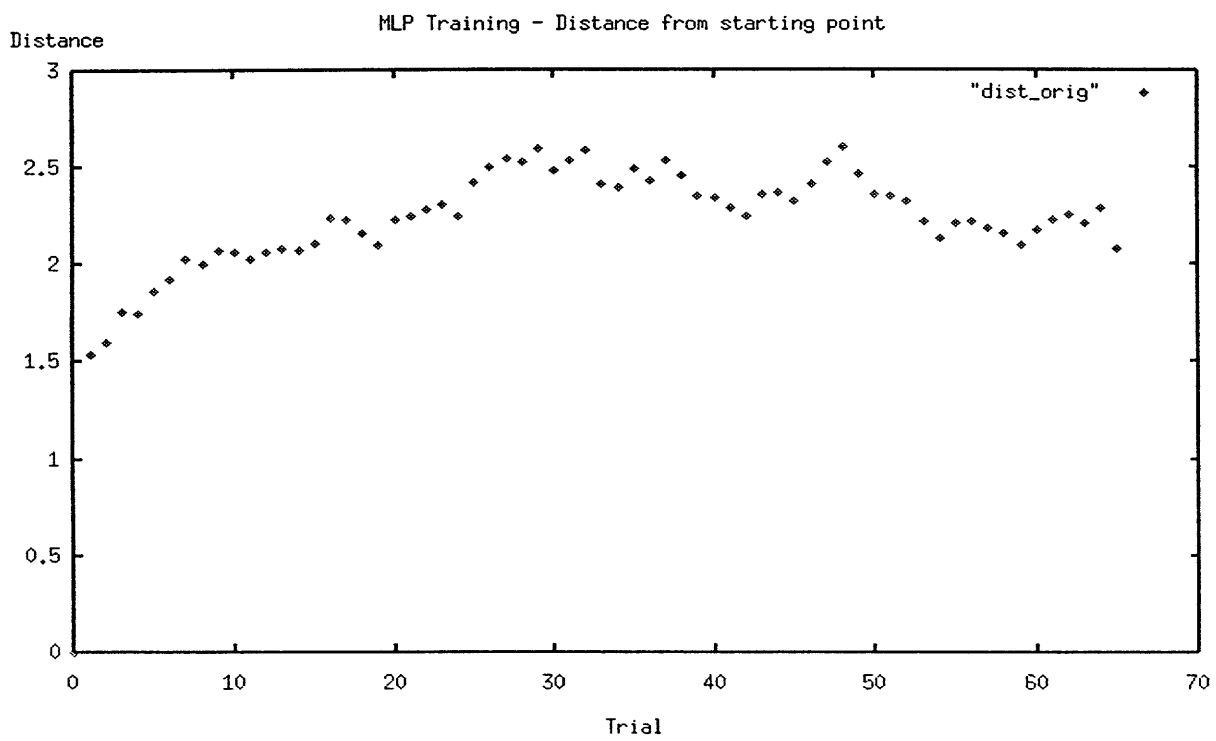


Figure 4-5: MLP Training - Distance from Starting Point

Point	TER (%)	Distances								
		A	B	C	D	E	F	G	H	I
A	11.839									
B	11.865	0.196								
C	11.865	0.178	0.345							
D	11.878	0.334	0.346	0.428						
E	11.878	0.388	0.371	0.453	0.468					
F	11.878	0.494	0.594	0.412	0.561	0.679				
G	11.905	0.480	0.532	0.491	0.387	0.532	0.375			
H	11.905	0.416	0.524	0.363	0.475	0.590	0.188	0.283		
I	11.905	1.178	1.286	1.084	1.262	1.374	0.932	1.151	0.961	
J	11.918	1.240	1.329	1.181	1.220	1.419	0.934	1.088	0.975	0.493

Table 4.6: Training results for the VLD using MLP for decisions, showing the ten points with best TER performance, out of 64 training runs.

point distances are shown in Table 4.6. Although the top three are fairly close to each other, there are examples of points with good performance that are fairly distant from each other (up to 10 or 15 times the iteration distance of 0.1). This supports the visualization of the error function as a broad valley, with a flat or very shallow bottom, but with many small bumps and hollows in it. A broad optimum for the error function is desirable because it means that the performance of the VLD will not be critically dependent on the exact values chosen for the parameters.

To examine the robustness of performance against small changes in parameter value, the point with the best TER performance (labeled A in Table 4.6) was used as a starting point. Eight new points were generated, each by annealing from A (with a smaller interval of 0.08) and retraining (with an initial step interval of 0.08). The performance of these points, and their distances from each other, are shown in Table 4.7.

All of the points yield performance fairly close to the base point, with some points being a bit higher and some a bit lower, and the points vary in distance from each other, indicating a “valley” with a fairly flat bottom, which is a good indication of stability and insensitivity

Point	TER (%)	Distances								
		Base	B	C	D	E	F	G	H	I
1	11.839	0.000								
2	11.984	0.195	0.195							
3	11.852	0.150	0.150	0.327						
4	11.826	0.240	0.240	0.320	0.288					
5	11.866	0.205	0.205	0.258	0.275	0.208				
6	11.747	0.090	0.090	0.243	0.139	0.219	0.213			
7	12.037	0.124	0.124	0.256	0.195	0.302	0.308	0.156		
8	11.800	0.199	0.199	0.335	0.246	0.262	0.222	0.227	0.234	
9	11.800	0.104	0.104	0.244	0.182	0.198	0.235	0.117	0.141	0.204

Table 4.7: Training results for the VLD using MLP for decisions, showing the results of nine annealing runs based on point A from Table 4.6.

to the exact values of the MLP weights.

Therefore, point number 6 from Table 4.7 was chosen as the final result. The network values of the resulting MLP are shown in Figure 4-6. Comparing these values to Figure 4-3, we see that most of the weights have changed somewhat, but not drastically. The only large changes are from Level to the lower hidden unit (-1.37 to -2.63) and the output bias (0.87 to -0.35).

4.6 Error characterization

With the final version of the VLD in hand, we wish to examine its performance in detail, with special attention to the circumstances in which it makes errors. The RVLM method for computing error rate (as described in section 4.2.2) is useful for training but does not provide much detailed information about the circumstances which lead to errors. As we will discuss in section 5.1.2, characterization of the phenomena which lead to VLD errors will be a vital part of subsequent processing using the Vowel landmarks.

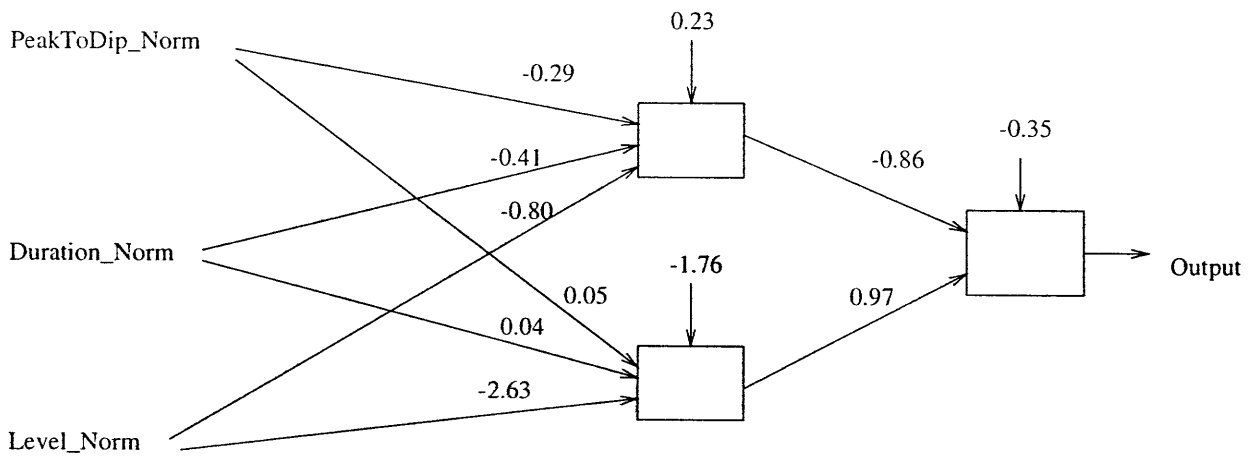
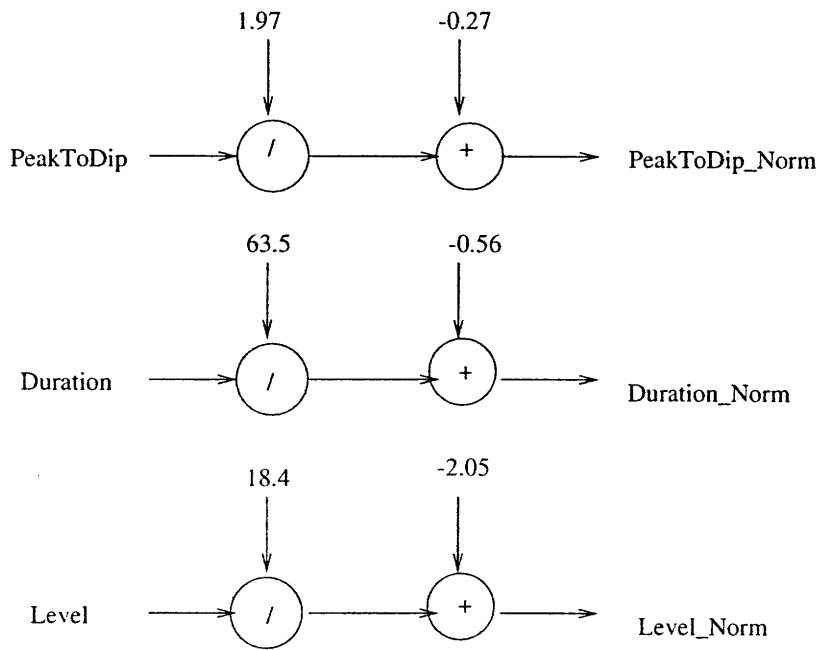


Figure 4-6: MLP network weights, final version. These weights result from training using gradient descent on an error rate criterion. See the text for interpretation of the values.

	Train		Test	
	Count	Percent	Count	Percent
Files	619		373	
Vowels	8104		4709	
Detections	6159	76.0	3556	75.5
Deletions	1945	24.0	1153	24.5
Insertions	1101	13.6	651	13.8
Error Rate	3046	37.6	1804	38.3

Table 4.8: Test results for the final VLD using canonical error categories. Percentages are relative to vowel count. Error rate is insertions plus deletions.

4.6.1 Canonical error categories

Statistics were gathered for the three main results of the VLD – Detection, Deletion and Insertion. With no additional qualifications, and a strictly literal interpretation of the TIMIT labels, this is the simplest measure of performance, but also the least accurate reflection of what we want the VLD to do.

The VLD, using the MLP with the network values chosen in section 4.5.4, was tested on the Core Training and Core Test sets (as described in section 4.2.3). The results are shown in Table 4.8. Results are very similar for the training and test sets, which indicates that the training set is large enough to capture essentially all the variation in the database.

The error rate (around 38%) is not impressive. Compare these values to the results shown in table 4.2, with error rates around 14%, or to the training results shown in table 4.7, with error rates below 12%. The difference is the previous results were generated using the RVLM method described in section 4.2.2. The main insight to be gained is that strict interpretation of the TIMIT labels is not a particularly accurate measure of the desired behavior of the VLD.

4.6.2 Modified error categories

To provide more insight into the VLD's behavior, the three main categories were divided into subcategories. The Insertion category was separated into insertions in consonants, semivowels, and vowels.

Deletions were separated into simple deletions and deletions in vowel-vowel context (VV), which were defined as a deleted vowel adjacent to a different, detected vowel. This latter condition corresponds to the optional Landmark capability of RVLMS which were used in training (see section 4.2.2).

Modification: Skewed detection

A special condition, called a skewed detection, was defined as a deleted vowel adjacent to a Vowel landmark inserted in a semivowel (that is, the semivowel with inserted landmark abuts the vowel with no landmark). Under this condition, the landmark is considered to be reasonably placed, while the boundary between vowel and semivowel is considered to be arbitrary and not reliable. Such boundaries were set by the TIMIT transcribers without regard to acoustic evidence, see section 1.3.2.

A skewed detection is considered to be a reasonable detection result; however, it is not just a subcategory of detection. Neither the consonant insertion nor the vowel deletion are counted as errors in scoring. (If this is not done, each skewed detection would correspond to not one but two errors. The resulting "double counting" of errors is liable to have a major impact on error statistics.)

	Train		Test	
	Count	Percent	Count	Percent
Files	619		373	
Vowels	8104		4709	
Detect, simple	6159	76.0	3556	75.5
Detect, skew	626	7.72	359	7.62
Delete, simple	900	11.1	560	11.9
Delete, VV	419	5.17	234	4.97
Insert, vowel	78	0.96	55	1.17
Insert, semivowel	30	0.37	17	0.36
Insert, consonant	367	4.53	220	4.67
Error Rate	1375	17.0	852	18.1

Table 4.9: Test results for the final VLD using modified error categories. Percentages are relative to vowel count. Error rate is all insertions plus simple deletions.

Results

The VLD, using the MLP with the network values chosen in section 4.5.4, was tested on the Core Training and Core Test sets (as described in section 4.2.3). The results are shown in Table 4.9. This is essentially a more detailed breakdown of the errors in Table 4.8.

The most significant observation is the fairly large number of skewed detections (7.7%). Because each skewed detection corresponds to two errors in Table 4.8, this accounts for 15% of the error rate. Deletions in VV sequences (about 5%) are less frequent but still make a difference. Together, these two modifications reduce the error rate to less than half the value from Table 4.8.

Essentially all the remaining insertions are in consonants (almost none in vowels or semivowels). The insertions in consonants deserve further investigation. Also, simple deletions constitute the single largest contribution to total error rate, and these deserve further investigation.

Manual examination of some examples seems to indicate an unexpected and rather surprising

phenomenon – skewed detections can happen when vowels adjoin obstruent consonants (as well as sonorants). Just as with semivowels, the boundary between consonant and vowel may be placed in a rather arbitrary way which does not properly represent the acoustic information. Usually, the vowel in question is destressed or reduced, and the consonant in question is a voiced fricative, glottal stop, or /h/.

An important example is the word "the" which appears frequently in the corpus. It is generally transcribed /dh ax/ with the boundary placed more or less in the middle of the corresponding sound. Just as with semivowels, a slight misalignment of the boundary causes two errors: deletion of the vowel and insertion in the adjoining consonant.

It is still the case that the error rates shown in table 4.9 (17% to 18%) are greater than in table 4.2 (around 14%), or in table 4.7 (around 12%), which were generated using the RVLm technique described in section 4.2.2. It is apparent that some cases which count as errors in table 4.9 do not count as errors when using RVLms for scoring. This will be explored in the next section.

Characteristics of detections

Table 4.10 shows the vowel detection statistics by category, using the same vowel categories as in table 3.1 (but not using RVLms). Diphthongs and lax vowels are the most often detected, while tense vowels are not detected as often. This may result from the phonotactics of tense vowels, which can appear in vowel-vowel contexts, while lax vowels cannot.¹ (Vowels which are deleted in vowel-vowel contexts are not counted among the detected vowels in Table 4.10, even though they are not included in the error rate either.) Schwas and syllabic sonorants are the least often detected, which is not surprising since they are generally reduced in amplitude

¹Except for lax vowels preceded by schwas, as in "they are interested."

	Train		Test	
	total vowels	percent detected	total vowels	percent detected
Total	8104	83.7	4709	83.1
schwa	2527	70.7	1483	68.9
sonorant	317	68.1	150	70.7
lax	2170	93.5	1220	93.7
tense	2628	87.6	1531	87.0
diphthong	462	97.6	325	96.0

Table 4.10: Category statistics of detected vowels, using the same vowel categories as in table 3.1.

	Train		Test	
	Count	Percent	Count	Percent
Total	475		292	
stop	195	41.1	127	43.5
fricative	119	25.1	64	21.9
affricate	8	1.68	5	1.71
nasal	23	4.84	10	3.42
semivowel	42	8.84	25	8.56
vowel	73	15.4	50	17.1
other	15	3.16	11	3.77

Table 4.11: Manner characteristics of segments with LM insertions.

and duration.

Characteristics of LM insertions

Table 4.11 shows the manner characteristics of segments with landmark insertions. Most are stops (about 42%) and fricatives (about 23%), although it is worth remembering that there are very few landmark insertions overall, as seen in Table 4.9.

Since a large proportion of landmark insertions occur in stops (at least, in regions which the TIMIT labeling denote as stops), we wish to examine the stops in more detail. Table 4.12

	Train		Test	
	Count	Percent	Count	Percent
Total	195		127	
b	3	1.54	5	3.94
d	21	10.8	10	7.87
dx	4	2.05	0	0
g	2	1.03	2	1.57
gcl	1	0.513	0	0
k	52	26.7	28	22.0
p	10	5.13	17	13.4
q	49	25.1	17	13.4
t	53	27.2	48	37.8

Table 4.12: Statistics of stops with LM insertions. The percentages are relative to the total number of LM insertions in stops.

shows the breakdown of landmark insertions in stops by label. Insertions are most frequent in segments labeled k, t, and glottal. The voiceless stops may well generate a burst of energy close to the F1 range (especially in a rounded context) which is more energetic than its surroundings. We observe that /p/ does not cause such frequent insertions, however. We may expect that a rounded context does not affect labial stops in the same way, since the source for burst and fricative energy is at the lips and not as strongly affected by rounding. Voiceless stops may cause more insertions than voiced stops because the burst is more isolated from surrounding regions of high energy at low frequencies, which makes it look more vowel-like to the VLD. The glottal stops, however, seem less likely to behave in this way, and are more likely to have labels that are placed inappropriately, leading to skewed detections, as described above (section 4.6.2).

Likewise, Table 4.13 shows the breakdown of landmark insertions in fricatives by label. Insertions are most frequent in segments labeled /dh/ and /s/. The voiced non-strident appears frequently in function words such as “the” which are prone to have the labels placed inappropriately, leading to skewed detections, as described above (section 4.6.2). The /s/ is frequently very loud, and may cause significant energy close to the F1 range (especially

	Train		Test	
	Count	Percent	Count	Percent
Total	119		64	
dh	52	43.7	24	37.5
f	3	2.52	1	1.56
s	38	31.9	26	40.6
sh	6	5.04	2	3.13
th	8	6.72	1	1.56
v	5	4.20	5	7.81
z	7	5.88	5	7.81

Table 4.13: Statistics of fricatives with LM insertions. The percentages are relative to the total number of LM insertions in fricatives.

	Train		Test	
	Count	Percent	Count	Percent
Total	1319		794	
schwa	741	56.2	461	58.1
sonorant	101	7.66	44	5.54
lax	140	10.6	77	9.70
tense	326	24.7	199	25.1
diphthong	11	0.834	13	1.64

Table 4.14: Vowel categories of deletion errors. using the same vowel categories as in table 3.1.

in a rounded context). This would make the /s/ appear vowel-like to the VLD if its low frequency energy is stronger than its surroundings (as it would if the surrounding segments were stops, as in “sixty” or the rounded context “looks to”).

Characteristics of vowel deletions

Table 4.14 shows the statistics of deleted vowels by vowel categories, using the same vowel categories as in table 3.1. Most are schwas (about 57%), and some tense (about 25%).

If subglottal pressure drops significantly at the beginning and end of the utterance, it is

reasonable to suppose that there might be more vowel deletions at the beginnings and ends of utterances than there are in the middle of the utterance. To test this hypothesis, the percentage of vowel deletions was computed as a function of the vowel's position in the utterance. Histograms of percent deletions, counting from the beginning and from the end of the sentence, are shown in Figure 4-7. However, neither histogram appears to show a higher percentage of deletions at the utterance boundaries than elsewhere in the utterance.

4.6.3 Additional modification: Skewed detections in consonants

As mentioned in section 4.6.2, it appears that skewed detections can happen when vowels adjoin obstruent consonants (as well as semivowels). From manual inspection of some examples, it appears that the vowel in question is usually distressed or reduced, and the consonant in question is usually a voiced stop, voiced fricative, glottal stop, or /h/, and both of the labeled segments in question are quite short (which may make their endpoints more difficult to place accurately).

A new error category was added for skewed detections in consonants other than semivowels. (The category of skewed detections in semivowels was left unchanged.) The statistical measurement of Table 4.9 was repeated using this new category.

The VLD, using the MLP with the network values chosen in section 4.5.4, was tested on the Core Training and Core Test sets (as described in section 4.2.3). The results are shown in Table 4.15.

The addition of skewed detections in consonants brings the count of insertions in consonants down from 367 to 158 (train) or from 220 to 93 (test), eliminating more than half the insertions in consonants. This is a substantial change, which indicates that many (if not

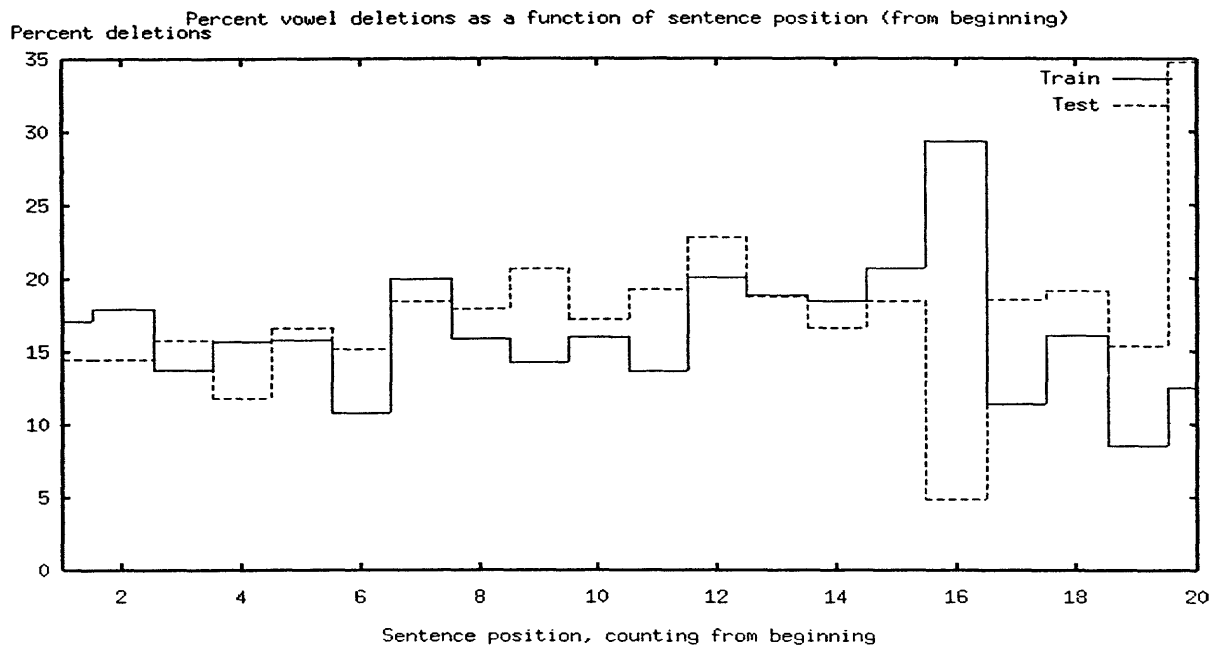


Figure 4-7: Histograms of percent vowel deletions as a function of sentence position. The horizontal axis is the vowel count, counting from the beginning of the sentence (upper) and from the end of the sentence (lower).

	Train		Test	
	Count	Percent	Count	Percent
Files	619		373	
Vowels	8104		4709	
Detect, simple	6159	76.0	3556	75.5
Detect, skew in semivowel	626	7.72	359	7.62
Detect, skew in consonant	209	2.58	127	2.70
Delete, simple	692	8.54	434	9.22
Delete, VV	418	5.16	233	4.95
Insert, vowel	78	0.96	55	1.17
Insert, semivowel	30	0.37	17	0.36
Insert, consonant	158	1.95	93	1.97
Error Rate	958	11.8	599	12.7

Table 4.15: Test results for the final VLD using modified error categories, including skewed detection in consonants. Percentages are relative to vowel count. Error rate is all insertions plus simple deletions.

most) of the insertions in consonants were due to poorly placed labels in the transcription.

The addition of skewed detections in consonants brings the count of (simple) vowel deletions down from 900 to 692 (train) or from 560 to 434 (test). This change is not as substantial as for insertions, since there are many more deletions than insertions in general.

By comparison, vowels which are detected when adjacent to other vowels are 558 or 6.88% (train), 333 or 7.07% (test), which is not very many more than the counts of vowels deleted when adjacent to other vowels (418 train, 233 test). This means that not very many pairs of abutting vowels have both vowels detected. Those that do constitute 140 or 1.72% (train), 100 or 2.12% (test), which is only about one eighth of all VV sequences.

The error rates in table 4.15 (around 12%) are about the same as the result of the MLP training using RVLMs, shown in table 4.7. This indicates that the skewed detections in consonants were causing the disparity between the error rates shown in table 4.9 and error rates based on RVLM scoring. We now have some confidence that the categories in table 4.15

	Train		Test	
	Count	Percent	Count	Percent
Total	209		127	
stop	112	53.6	69	54.3
fricative	71	34.0	43	33.9
affricate	4	1.91	4	3.15
nasal	17	8.13	9	7.09
semivowel	3	1.44	1	0.787
vowel	0	0	0	0
other	2	0.957	1	0.787

Table 4.16: Manner characteristics of LMs for skewed detections in consonants.

	Train		Test	
	Count	Percent	Count	Percent
Total	209		127	
schwa	173	82.8	113	89.0
sonorant	3	1.44	2	1.57
lax	10	4.78	5	3.94
tense	19	9.09	7	5.51
diphthong	4	1.91	0	0

Table 4.17: Vowel categories of vowels which show skewed detections in consonants, using the same vowel categories as in table 3.1.

encompass all the phenomena which are subsumed in the RVLM scoring technique, and show their relative frequencies.

Table 4.16 shows the manner characteristics of consonants which cause skewed detections. Almost all are stops and fricatives. Table 4.17 shows the categories of vowels which cause skewed detections. Almost all are schwas. Both of these results confirm the impressions resulting from manual inspection in section 4.6.2.

Of the all the skewed detections in consonants, almost all are “backward” skews, where the consonant precedes the vowel (199 out of 209 train, 115 out of 127 test).

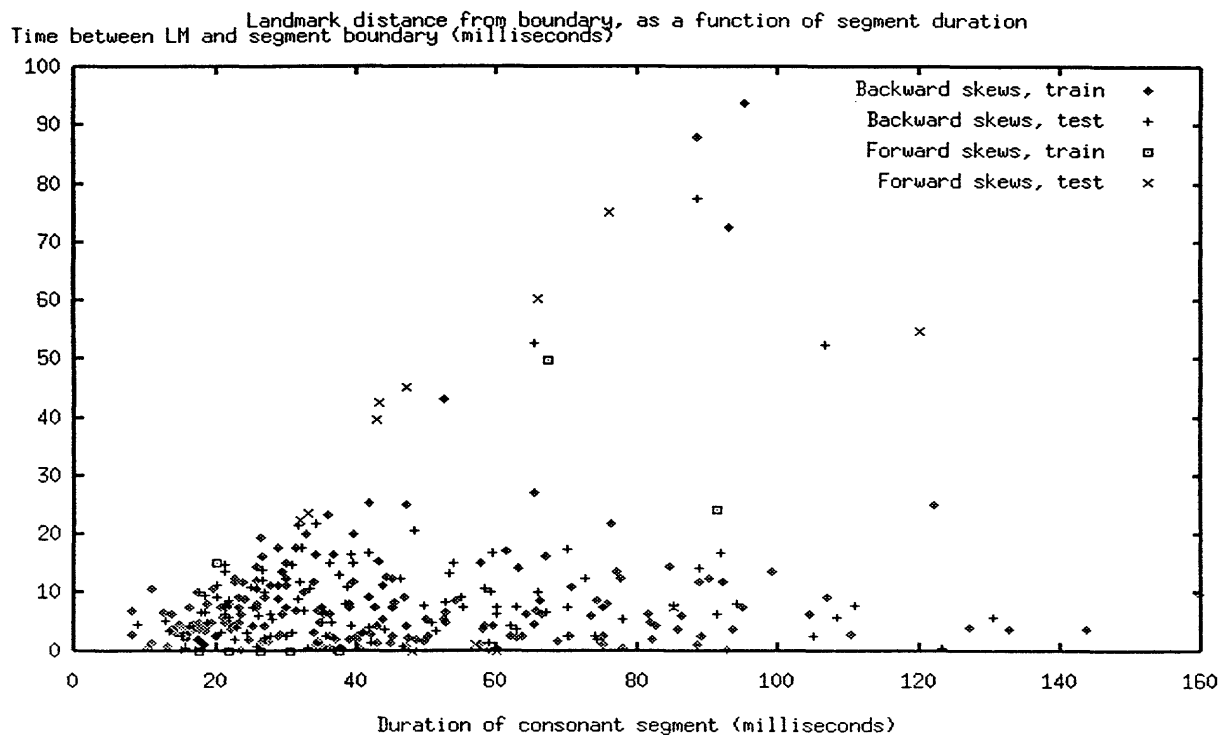


Figure 4-8: Scatter plot of skewed detections in consonants, plotting the time between the landmark and the segment boundary against the duration of the consonant segment. Most of the segments are fairly short, and most of the landmarks are close to the segment boundary.

For most skewed detections in consonants, we expect the landmark to be fairly close to the boundary between the consonant and the vowel (because the skewing results from inaccurate placement of the boundary during labeling). Figure 4-8 shows a scatter plot of skewed detections in consonants, plotting the time between the landmark and the segment boundary against the duration of the consonant segment. Most of the skewed detections in consonants occur in rather short segments, and most of the landmarks are quite close to the segment boundary, even for long segments.

4.6.4 Conclusions

The primary conclusion of section 4.6 is that scoring with the RVLM technique (section 4.2.2) is quite different from scoring with more conventional measures (as in table 4.8). The differences are due to skewed detections, which are really artifacts of the segmental labeling of the TIMIT database, and deletions of vowels in vowel-vowel sequences.

Strict, conventional scoring yields error rates about 38%. Deletions in vowel-vowel sequences account for about 5%, skewed detections in semivowels account for about 15%, and skewed detections in consonants account for about 4% or so (see table 4.15). When these phenomena are taken into account, the resulting error rate is just a bit below 12%, matching the error rate measured with RVLMS.

4.7 Examples from the TIMIT database

This section presents examples of utterances from the TIMIT database with landmarks generated by the VLD. The Vowel landmarks are labeled with the hybrid confidence score described in section 5.2.3. It is clear that most of the Vowel landmark decisions are made by the hard limits, as indicated by the unity confidence score values.

The first example demonstrates phenomena in vowel-semivowel clusters, and appears in figure 4-9. From top to bottom, the panes show the waveform, wide-band spectrogram, low frequency energy track, TIMIT labeling (both aligned phones and words), and Vowel landmarks.

There are instances of skewed detections at 0.64 s (“you”) and at 1.04 s (“we”). In each case,

the landmark occurs in the semivowel preceding the vowel. There is also a more unusual instance of skewing at 1.28 s and 1.42 s (the first two vowels in “were away”). Here both landmarks appear in the segment labeled /er/. The first is right next to the boundary of the preceding /w/ and the second is right next to the boundary of the following /ax/. Under the current scoring scheme, this is counted as a vowel insertion followed by a VV deletion. Despite the effort to account for labeling artifacts, this example shows a case where at least one error (the insertion) occurs as a result of arbitrarily placed labels, even though the VLD is not doing anything obviously wrong.

The second example demonstrates phenomena in vowel-vowel clusters, and appears in figures 4-10 and 4-11. From top to bottom, the panes show the waveform, wide-band spectrogram, low frequency energy track, TIMIT labeling (both aligned phones and words), and Vowel landmarks.

There are instances of VV deletions at 0.60 s (the second vowel in “triumph-”) and at 2.65 s (the first vowel in “heroism”). In each case the deleted vowel appears as a shoulder on the adjacent vowel. There is a more complicated example at 1.40 s, where the final vowel in “warrior” has deleted vowels on both sides. As the labeled sequence is /iy er ih/, vowels in these circumstances are not always labeled as reduced. The procedure which detects VV deletions allows both of these vowels to be labeled as VV deletions.

In contrast, correct detection of two vowels in sequence occurs at 2.30 s (“naive”) and correct detection of an epenthetic insertion of a vowel occurs at 3.00 s (final nasal in “heroism”). It is also worth noting that vowels separated by obstruent consonants are detected correctly, even when the vowels are very short, as in the three short vowels in a row around 1.80 s (“exhibited”).

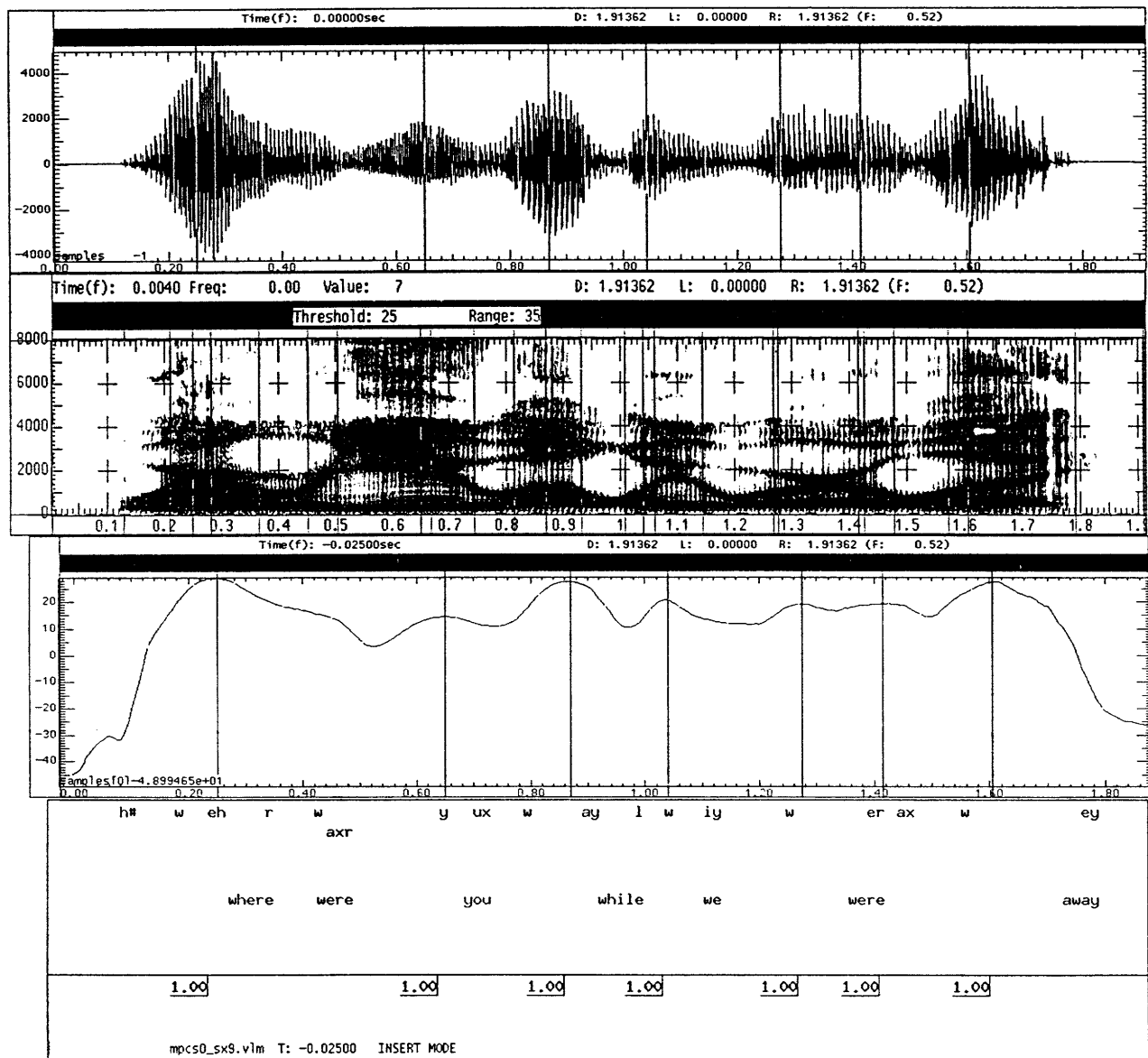


Figure 4-9: TIMIT Vowel-semivowel example. The sentence is SX9 “Where were you while we were away” uttered by male talker PCS0. Skewed detections occur at 0.64 s (“you”) and at 1.04 s (“we”), and two-sided skewing occurs at 1.28 s and 1.42 s (first two vowels in “were away”).

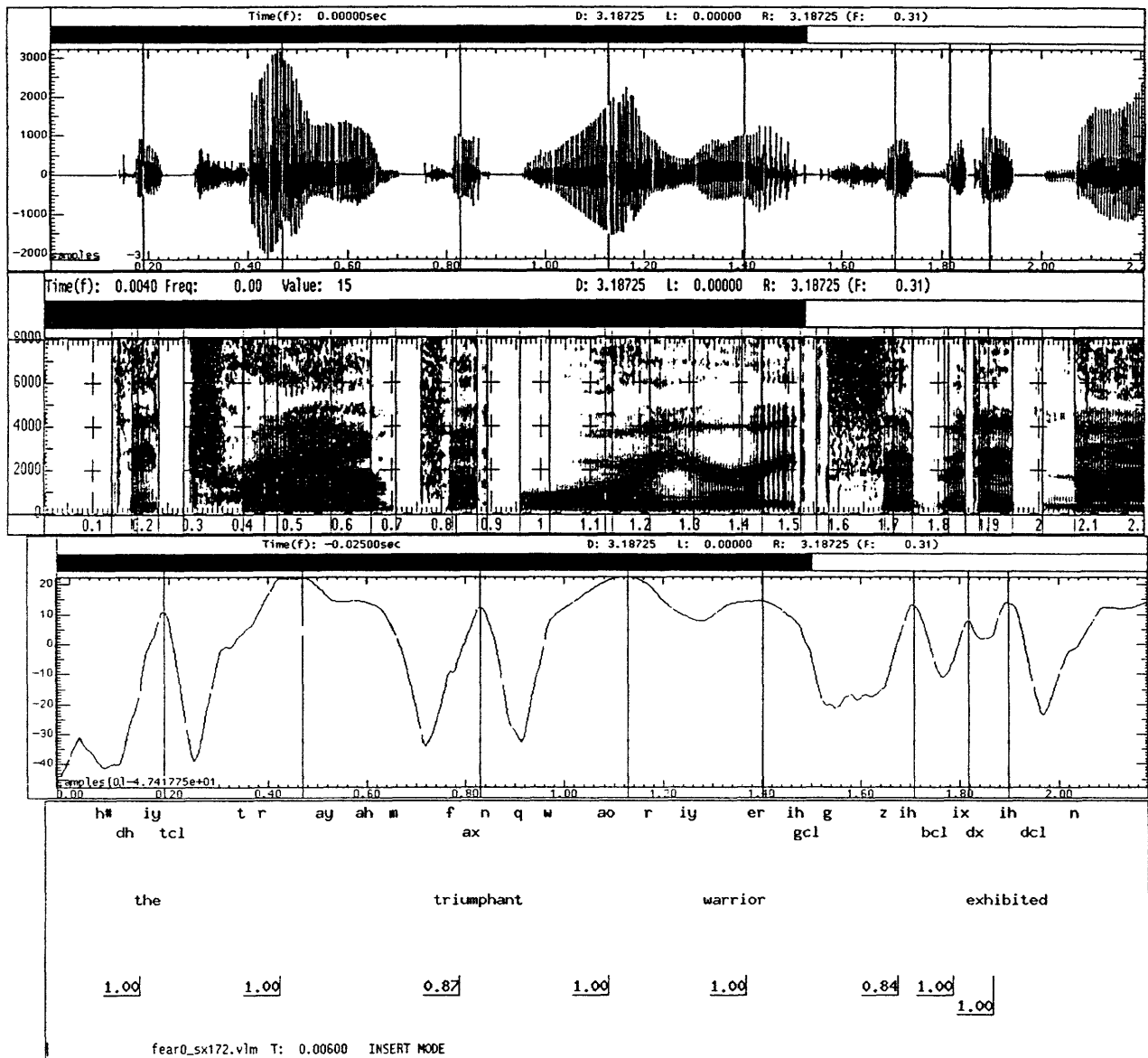


Figure 4-10: TIMIT Vowel-vowel example (page 1). The sentence is SX172 “The triumphant warrior exhibited naive heroism” uttered by female talker EAR0. VV deletions occur at 0.60 s (the second vowel in “triumph-”) and 1.40 s (the second vowel in “warrior” and the first vowel of “exhibited”), each appears as a shoulder on the adjacent vowel.

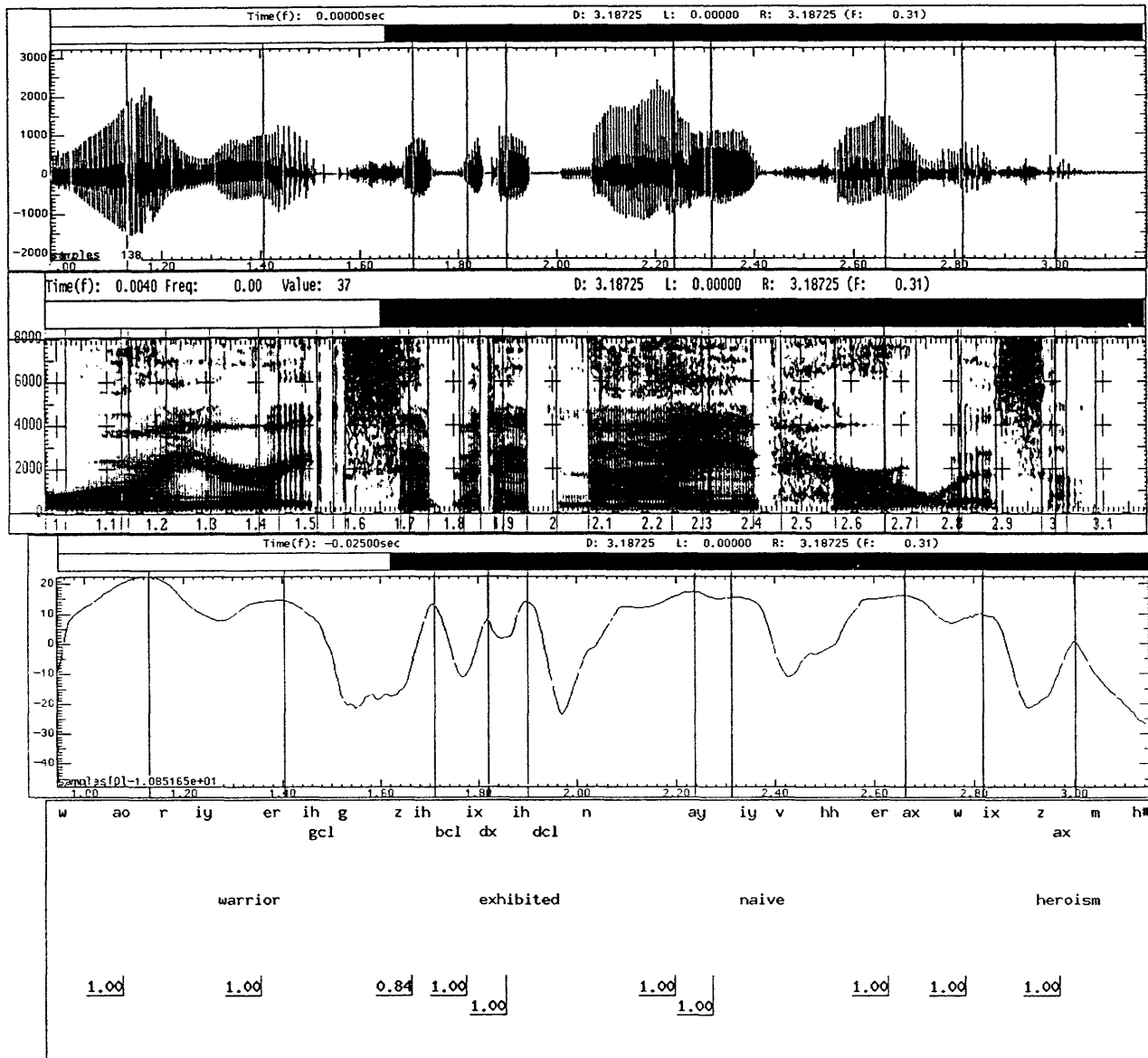


Figure 4-11: TIMIT Vowel-vowel example (page 2). The sentence is SX172 “The triumphant warrior exhibited naive heroism” uttered by female talker EAR0. Correct detection of two vowels in sequence occurs at 2.30 s (“naive”) and epenthetic insertion of a vowel occurs at 3.00 s (final nasal in “heroism”).

4.8 Conclusions

As seen in Table 4.15, the final version of the VLD yields a fairly good error rate (11.8% train, 12.7% test). Of these errors, most were deletions (8.54% train, 9.22% test) and few were insertions (3.28% train, 3.50% test). We recall from the experiments in Chapter 3 (Table 3.5) that 94.2% of all vowels in the TIMIT database show a proper peak in F1 amplitude, which implies that a VLD using peak picking on F1 amplitude cannot do better than 5.8% deletions. The VLD presented here comes reasonably close to this ideal.

In addition to the VLD itself, two other innovations have been presented in this chapter. The method for scoring the VLD by using reference Vowel landmarks (section 4.2.2) and the use of a neural network to combine knowledge-based acoustic measurements (section 4.5) are novel inventions, to the best of this author's knowledge. This author believes that these innovations are valuable, both for this thesis and for further work on the LAFF project.

Chapter 5

Implications and Future Work

5.1 Enhancements to Vowel Landmark detection

This thesis presents an algorithm which achieves good detection performance while maintaining fairly simple processing. However, it does not exhaust the possibilities for how to go about detecting Vowel landmarks.

5.1.1 Further improvements for VLD

There are more features that could be included in a Vowel Landmark detector. A measure of glottal excitation or voicing is one example, and a measure of the presence of formant structure (such as Hermes' spectral "peakiness" measure [37]) is another. Most such measures would be more complicated than the low frequency energy implemented in this thesis, and it is uncertain how much performance they would add.

Several researchers ([88], [41], [93]) have used absence of high frequency energy as a cue for syllabic nuclei (apparently to help distinguish them from frication). However, given the low frequency energy measure used in the present algorithm, it is doubtful that fricative regions present much of a problem.

There are other structures and algorithms that could be used for making the detection decision. There are potential gains to be realized here. For example, the convex hull recursion cannot detect a vowel which does not generate a peak in low frequency energy. Such vowels may happen fairly often in vowel-vowel sequences, where one vowel appears as a “shoulder” on another. The study of section 3.5.2 showed that about 6% of vowels do not show an amplitude peak around F1. Other techniques, such as dynamic neural networks or Hidden Markov Modeling, could be applied to this problem. Such techniques would be substantially more complicated than the convex hull recursion, however, and the potential gain in performance is rather small.

In sum, further work on the VLD itself would probably involve a large increase in complexity, and could only provide a small increase in performance.

5.1.2 Error characterization

Characterization of the errors made by the VLD will be very important for future work on a LAFF prototype system (which will be discussed more below, in section 5.4). The statistical study performed in section 4.6 is a good start in providing this kind of information, but there are more questions to be answered.

For example, the lexical access module of a LAFF system will need to know how often vowels are deleted in function words as opposed to content words, how errors depend on prosodic

context, and so on. In general, we will want to provide as much information as possible about what kind of vowels are most likely to be deleted, and what kind of phenomena are most likely to cause insertion errors. This information will be used in lexical access (see section 5.3.2 below).

5.1.3 Adaptability to other databases

As mentioned in section 1.4.2, one design goal for the VLD is ability to accept speech with a variety of characteristics such as talking rate, gender and dialect of talker, and talking style. The VLD should be insensitive to such variation or adapt to the variation as necessary. The TIMIT database encompasses a variety of talkers of American English, but there are other databases that may be used for further exploration.

The LAFF database [10] is a database under development at the Speech Communication Group. It is a database of read sentences recorded in quiet, by four talkers (two male, two female) with aligned landmark transcriptions.

The Switchboard database [28] is a database of spontaneous speech (two sided telephone conversations) of male and female talkers, from all areas of the United States. The database as published includes orthographic transcriptions and time aligned word transcriptions. The Speech Communications Group has some time aligned phonetic transcriptions for a portion of the Switchboard database, which were generated by Greenberg et al. [34].

Spontaneous speech in particular would be worthwhile to explore, as it is liable to manifest more production variability (see section 1.3.1) than the read speech of the TIMIT database. Epenthesis, elision, and coalescence are all liable to be more evident in spontaneous speech, particularly across word boundaries. Spontaneous speech is also subject to errors such as

partial words, hesitations and pauses, and filler sounds such as “um” which tend not to occur in read speech.

5.1.4 Vowel classification schemes

Presumably, in a LAFF system, Vowel landmark detection will be followed by processing to characterize the vowel quality, or the features of the vowel (primarily the features high, low, back, tense/lax, and round). The task of vowel classification has been studied by a number of researchers, and a variety of schemes have been tried. In general, formant frequencies are the primary acoustic cues to vowel class, whether chosen at a single instant or via a weighted average across the vowel duration [18]. Classification algorithms include analysis by synthesis [8], spectral similarity (as in the K nearest neighbors experiment of section 3.7.2), and so on. It may also be necessary to know at least some features of the surrounding consonants or other contextual information.

Characterizing vowel-vowel and vowel-sonorant sequences is liable to be more difficult. As we have seen in section 4.2.2, only one Vowel landmark is likely to appear in such regions. A scheme must be devised to allow for the possibility of more than one lexical vowel per Vowel landmark (for vowel-vowel sequences) or for concentration of vowel information in a different place from the Vowel landmark (for vowel-sonorant sequences). Perhaps such a scheme will postulate new Vowel landmarks (attempting to undo deletion errors) or remove existing ones (attempting to undo insertion errors).

Presumably, each landmark output by the VLD will serve as a starting point for vowel place or quality analysis, which will extract features such as high, low, back, tense, and so on. Vowel quality is notoriously difficult to measure. It may be that the vowel quality analysis will work best if the vowel landmark is near the center of the vowel (where “center” may

mean either its midpoint in duration, or its point of maximum opening). At this time, it does not appear that we can guarantee that vowel landmarks will be placed close to the vowel center. If the landmark location is important, we may want to investigate a procedure for “fine tuning” the location of the vowel landmark in time for this purpose. However, the experiment of section 3.7 indicates that location is not critical, so it may be that no fine tuning will be necessary.

5.2 Confidence scores

As discussed in section 1.3.4, the generation of confidence scores is important for the proper functioning of subsequent stages of processing, when the LAFF system is assembled. There are two parts to the implementation of confidence scores. First, we wish to generate a confidence score value which agrees with intuitive judgements. Second, we wish to validate that the confidence score carries meaningful information, or at least to demonstrate that higher scores are associated with higher probabilities of correctness.

5.2.1 Generation of confidence scores

Ideally, the VLD would generate a value suitable for use as a confidence score as part of its normal operation. Unfortunately, the implementation described in Chapter 4 does not generate a suitable value. The MLP output is used to make the decision to terminate the convex hull recursion. As such, it represents a judgement of the interpeak dip, not the peak itself (see figure 4-1 and the accompanying discussion for more details). Therefore its values do not always reflect intuitive judgements of the robustness of the peak.

In addition, the MLP output is not the only value which can terminate the convex hull recursion. As described in section 4.5.2, there are hard limits on the values of the acoustic measurements (peak-to-dip ratio, duration, and level) which operate separately from the MLP. Landmark decisions made by the MLP can easily be given a confidence score which is the output value of the MLP. But how the landmark decisions made by hard limits should be scored, in order to combine them with the “soft limited” MLP decisions, is not entirely clear.

There are several possibilities for other ways to compute a confidence score. A separate MLP could be added to the VLD, and trained to combine the various attributes of the peak (absolute level, peak-to-dip difference, and dip-to-dip duration) into a single confidence score. Other possibilities involve modifications to the convex hull algorithm. Simply inverting the low frequency energy track and running the same convex hull algorithm would result in decision values for peaks instead of dips. There may be alternative techniques which would avoid the use of hard limits, or entirely different algorithms for peak picking, such as the constrained techniques of Hermes [37] or Pfitzinger et al. [69].

In any case, the confidence score should take values between 0.0 and 1.0, and should be interpretable as the probability that the landmark represents an actual underlying vowel.

5.2.2 Validation of confidence scores

Intuitive judgement is not enough to validate a confidence score. It ought to carry objective information about the likelihood of the landmark’s correctness, interpretable in probabilistic terms. One method for validation is presented by Chun [13, p. 31], in which a histogram of detections by confidence score is computed (which Chun calls the histogram of probability estimates). This is really two histograms superimposed, one for “Correct”

decisions, and another for “All” decisions. “Correct” decisions include detections (which contribute to the bin at their confidence score) and insertions (which contribute to the bin at 1.0 - confidence score). “All” decisions include all landmarks, at both their confidence score bin, and 1.0 - confidence score bin.

The ratio of “Correct” decisions to “All” decisions is called by Chun the probability estimate ratio, and is likewise plotted as a function of confidence score. Ideally, this histogram should show a straight line rising from 0 to 1, for a perfect estimator. Deviations from a straight line are acceptable, if the line is monotonically increasing, because the shape can be adjusted by a warping function.

One problem is that confidence scores are only available for Vowel landmarks as they appear in the VLD’s output, which means that we can compare detections to insertion errors, but not deletion errors. Recall from section 4.6 that most of the errors made by the VLD are deletion errors, and insertion errors are relatively few. Therefore, there will not be very much data on which to base the measurements.

5.2.3 Example: Hybrid confidence score

As an initial investigation into the issues around confidence scoring, a hybrid score was created, using the MLP output for landmarks generated by MLP decisions, and the fixed value 1.0 for landmarks generated by hard limits. The MLP output value was scaled slightly to extend the data to cover the entire range from 0.0 to 1.0, and the few values which were slightly negative were clamped to zero.

For this hybrid confidence score, a histogram of probability estimates (landmark counts by confidence score) was computed. For these data, the hard-limit point at 1.0 constitutes over

half the entire data set (41207 out of 69044 detected landmarks, or 59.7%, and 45765 out of 76806 total landmarks, or 59.6%). It was evident that the hybrid confidence score did not generate an even distribution of values. This may result from the inadequacies of the MLP output as a confidence score, as described in the previous section.

A histogram of probability estimate ratio by confidence score was also computed. It did not show a monotonically increasing line at all points. Several definite deviations from monotonicity were apparent. The main insight to be gained from this exercise is that the hybrid confidence score is not a satisfactory estimator. One or more of the schemes described in section 5.2.1 may be used to derive a better confidence score.

5.3 System integration issues

The Vowel Landmark Detector presented in this thesis was developed in isolation. To be useful in a LAFF system, it must work well with subsequent stages of processing.

5.3.1 Optimization criteria

The Vowel Landmark Detector presented in this thesis was optimized for the goal of minimum error rate (where error rate is defined as in section 4.2.2). The appropriateness of this goal will depend on how the VLD is used in a LAFF system. As described in section 1.4.3, the goal of maximum confidence in Vowel landmarks may be achieved by treating insertion errors as more costly than deletion errors, while the goal of maximum information output may be achieved by treating deletion errors as more costly than insertion errors.

One advantage of the decision algorithm chosen for the final version of the VLD (section 4.5) is that the output is a single scalar value. Training the MLP to minimize error rate was done with an implicit threshold of zero (positive values indicate a Vowel landmark, negative values indicate no landmark). This threshold can be changed to adjust the error bias (a higher value will produce fewer insertion errors and more deletions, while a lower value will produce more insertions and fewer deletions) and the adjustment does not require retraining the MLP weights.

Although the final threshold value is easily adjustable, we do not know what value to choose in order to achieve a desired bias, or a desired probability of a given type of error.¹ In other words, we have a control, but we do not have a scale for it. Calibration of control parameters will make the VLD more usable in practical situations, and should be addressed. Calibration of output values has been discussed in section 5.2, and will be discussed further in section 5.4.3.

5.3.2 Lexical contact and error recovery

In Section 5.1.2 we discussed error prediction: characterizing which vowels are likely to be missed, and which phenomena are likely to generate false alarms. In order to use the Vowel landmarks to make contact with the lexicon, we need a procedure to recover from errors made by the VLD.

Some error recovery can be done by combining the Vowel landmarks with Consonant landmarks (which are generated by a separate process that does not use or look for Vowel landmarks, as in Liu [54]) and perhaps with Glide landmarks.² When the different classes

¹This may depend on the characteristics of error recovery in later stages of processing. For example, it may be more important to detect full vowels, or prominent vowels.

²Sun's method for finding Glide landmarks [86] depends on having the Vowel and Consonant landmarks

of landmarks are combined, they may be parsed into a basic syllable structure, using phonotactic constraints, durations, and other acoustic measurements to confirm landmarks that make sense and reject landmarks that do not. Syllable parsing from phonetic symbols has been studied before [42] (and implemented in [26]), but syllable parsing from landmarks is liable to raise different issues and problems. It may also be desirable to extract independent measurements from the acoustics to assist this process, such as estimates of sonorant and continuant features, as does Bitar [6].

In order to match items in the lexicon (which are strings of phonemes), the system must postulate segments from the landmarks. This process will be assisted by the syllable structure described above, but is almost certain to produce errors (substitutions, insertions, and deletions) if the landmark stream includes errors. The postulated segments should receive confidence scores, based on some combination of the landmark confidence scores and the feature confidence scores.

The lexical matching process itself has not received much attention to date (some aspects have been explored by Zhang [92]). A simple left-to-right matching technique seems the most straightforward, but it may not be sufficient in the event of errors. Left-to-right matching will need to allow for errors, perhaps by using a multipath state network to represent segments, including transitions that skip states (deletions) and including extra states (insertions), similar to the networks generated by DeMori [16]. Other strategies such as matching the high-confidence segments to stressed syllables first, which may be regarded as “islands of reliability” may be used, but this stage of processing will require substantial study.

Presumably, the result of the lexical matching process will be a group of words that are candidates to match the acoustic evidence, similar to the “cohort” of Marslen-Wilson [57]. Additional acoustic evidence will be sought to confirm or discard members of the cohort. The

available, and it is uncertain how much independent information his Glide landmarks carry.

search for additional acoustic evidence may include a search for additional Vowel landmarks. Such a search may be done by simply lowering the threshold of the output decision (at least in specified regions of the speech signal), or perhaps more complex changes will be required.

In contrast to the first pass described above (postulate segments from landmarks, and match segments to the lexicon), this second pass will involve synthesizing landmarks from candidate words in the cohort, and matching these landmarks to the acoustically derived information. Synthesized landmarks may include a value for how likely they are to be deleted, or to lack a corresponding acoustic event, as discussed in section 1.2.1, in which case the second pass can use this information during the matching procedure. This dual-mode matching is one of the hallmarks of the LAFF paradigm, and should be a crucial part of achieving good performance.

5.4 System design issues

There are more general issues that must be addressed in order to design and build a prototype LAFF system. Speech recognition algorithms require a strict, probabilistic formulation in order to use statistical and stochastic tools. The structures and rules of linguistic representation, however, do not generally include probabilistic information or allowance for conditions of uncertainty. Bridging this gap is one of the central problems of designing a LAFF prototype.

5.4.1 Acoustic cues and Information content

Each estimate of a Distinctive Feature (DF) will require measurement of some number of acoustic cues. The measurements of the acoustic cues must then be combined in some way to produce an estimate of the DF. In general, the combination of acoustic cues may not be straightforward. It may require nonlinear transformation, or there may be dependencies among cues (importance of one cue dependent on the value of another), or there may be trading relations between cues.

We will want a technique which is general and flexible enough to deal with complicated combinations. A Multi-Layer Perceptron (MLP) such as in section 4.5 has proven to be a useful tool in this situation, and will probably be useful for DF estimation as well.

Tradeoffs between acoustic cues

Additional problems arise when a single acoustic cue contributes to more than one DF. For example, voice onset time (VOT) is a cue for Voicing in stop consonants, but is also related to Place in stop consonants.³ If Voicing and Place are estimated independently, information will be lost. (For example, a VOT which is long because the stop is Voiceless may be misinterpreted as an indicator of velar Place, if the Place estimator is unaware of the Voiceless status.)

This problem can be solved by estimating Voicing and Place together, in one module. However, if there are many acoustic cues which contribute to more than one DF, then many modules would have to be combined, and the resulting system would not embody the or-

³Strictly speaking, it is the length of the frication burst that is a cue for Place in stops.

thogonality which makes distinctive features such a powerful concept in linguistics. If Voicing and Place are estimated sequentially (first one, and then the other), the second estimate will have the knowledge of the results of the first estimate, but the first estimate will have to be made “blind” which is what we want to avoid.

There are probably several different ways to deal with this problem. One possibility is to make an *a priori* assumption about the state of one DF in order to estimate another, and then combine the results of all possible assumptions. For example, the Place detector could output two results, one assuming a Voiced environment, and the other assuming a Voiceless environment. Independently, the Voicing detector could output one result for each presumed value of Place. Then a combination procedure would be used, which compares the output values of the two detectors and produces a final value for each DF. This procedure would probably need to use confidence estimates as well as DF values (see section 5.4.3 below for discussion of confidence estimates).

5.4.2 Representation of Distinctive Features

In general, lexical DFs are binary (which actually means they can take three values, +, -, and unspecified), although some are unitary (taking only two values, + and unspecified). Acoustic measurements, however, can take on a range of values, depending both on the presence or absence of acoustic information and what that information indicates about the value of the underlying feature. A LAFF system will require some method of representation of DFs that corresponds to lexical values, while maintaining as much information about the acoustic evidence as possible. (As discussed in section 1.3.4, maintaining information and avoiding “hard” decisions is essential to avoiding cascade failure.)

This section proposes that DFs derived from acoustic measurements should result in two

numbers: the Value (floating point between -1.0 and +1.0) which indicates what the acoustic information says about the underlying DF, and the Confidence (floating point between 0.0 and 1.0) which indicates the reliability of the acoustic information. These numbers will usually be given as an ordered pair (Value, Confidence).

A scheme for translation between these two representations is shown graphically in figure 5-1. The acoustic Value is plotted on the vertical axis, and the acoustic Confidence is plotted on the horizontal axis.

Translation of DFs from lexical representation to acoustic derivation is straightforward. A lexical + becomes an acoustic (+1.0, 1.0), and a lexical - becomes an acoustic (-1.0, 1.0) as indicated by the small circles.

Translation of DFs from acoustic derivation to lexical representation is a bit less clear. This document proposes that the acoustic space be divided into three regions (indicated by the dashed lines) so that acoustic values in the "+" region are translated to lexical +, acoustic values in the "-" region are translated to lexical -, and the remainder are translated to lexical "unspecified."

The exact shape of the region boundaries is not certain. Perhaps an exponential curve would be reasonable, with a decay coefficient to be determined.

5.4.3 Calibration of Feature Values

For the scheme in section 5.4.2 to be useful, the Value and Confidence scales need to be calibrated somehow, so that the outputs of different modules can be compared to each other. The shape of the region boundaries in figure 5-1 will probably depend on this calibration.

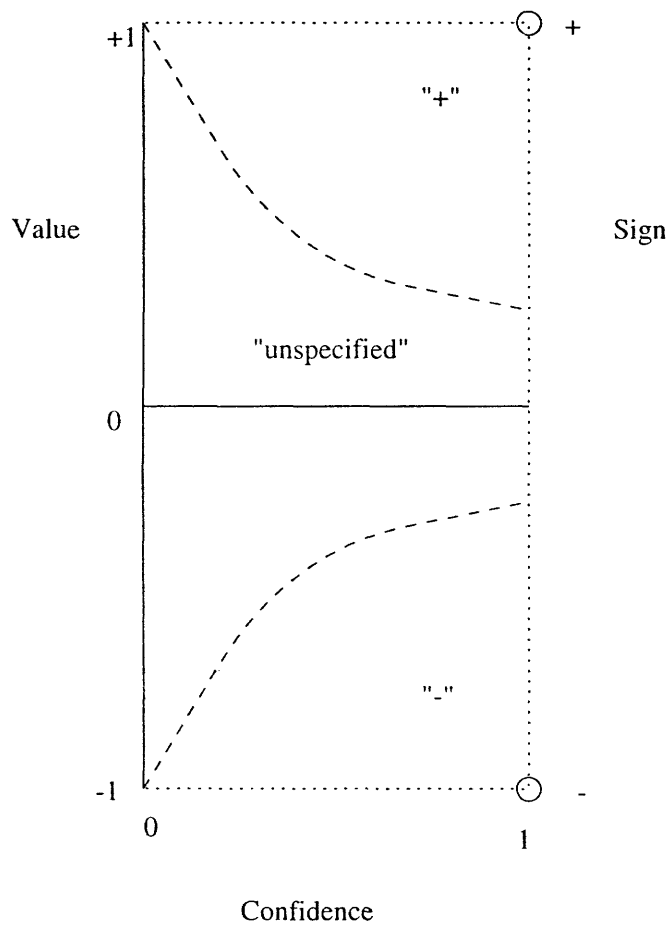


Figure 5-1: Translation scheme for Distinctive Features, showing how continuous values (derived from acoustics) relate to discrete values (represented in the lexicon). The acoustic Value is plotted on the vertical axis, and the acoustic Confidence is plotted on the horizontal axis.

It is uncertain what kind of calibration scheme would work best. A probabilistic interpretation seems appropriate, in which the output value reflects the statistical likelihood of the measurement being correct on a labeled database of speech. Validation of landmark confidence scores has already been explored somewhat in section 5.2. However, a strict definition of meaning of the confidence score and the details of a procedure to derive it are uncertain, and deserve further study.

5.4.4 Landmark and Feature Hierarchy

In the author's understanding of the current LAFF scheme [80], acoustically derived DFs are attached to Landmarks. A Landmark does not have a Value (as in section 5.4.2) but it should have a Confidence score. The DFs attached to a Landmark will have their own Confidence scores, whose computation will reflect the confidence of the Landmark. Probably the DF Confidence should be understood to be predicated on the Landmark Confidence – that is, the DF Confidence should represent the probability that the Value is correct, given that the Landmark is correct. Then, the total confidence in the DF is the product of the Landmark Confidence and the DF Confidence.

Likewise, if the DFs are organized in a feature hierarchy [43], and if there are DFs at intermediate levels of the hierarchy, the Confidence scores of child DFs may be predicated on the Confidence scores of their parent DFs (as well as the Confidence of the Landmark). In particular, most articulator bound DFs (e. g. distributed, anterior, lateral) will be predicated on the confidence of the Place feature which specifies their articulator (tongue blade).

5.4.5 Lexical Matching and Phonetic Rules

For lexical matching, the rules of DF transformation will also need to be phrased in probabilistic terms. For example, we know that a vowel which follows a coronal consonant will tend to be fronted. We will need to have quantitative data on how fronted it will appear, and how much variation can be expected, and how this transformation depends on other environment variables such as stress, prosody, rate of speech, and so on.

The lexical matching procedure itself has not received much attention. For the first “bottom-up” pass, segments may be hypothesized from the Landmarks, and matched to the segments of the lexicon. Based on this partial information, additional Landmarks and DFs may be hypothesized from lexical items, and a second “top-down” pass may be performed, looking for evidence of the hypothesized Landmarks and DFs. In either case, there need to be mechanisms in place to deal with landmark errors, both insertions and deletion, as described in section 5.3.2. The details of this process are not clear, and deserve further study.

Top-down rules could be generated to predict landmark errors from context. For instance, a vowel (especially a diphthong) followed by a liquid may generate an extra vowel landmark, as in “fire” or “feel,” unstressed vowels may be deleted, and so forth, as described in section 1.3.1. There may also be substitution errors, where glide landmarks correspond to vowel segments and vice versa.

Bottom-up error correction depends primarily on confidence scores, as described in section 1.4.2. By maintaining confidence scores, perhaps with lowered thresholds, more information is passed from the VLD to following stages of processing. The system can post process the landmark stream with phonotactic rules, which will help identify errors and correct them.

Bibliography

- [1] Barnett, Jeffrey A. et al. "The SDC Speech Understanding System." in [48], pp. 272-293.
- [2] Bishop, Christopher M. *Neural networks for pattern recognition*. Oxford University Press. 1995.
- [3] Bitar, N. and Espy-Wilson, C. Y. "A Signal Representation of Speech Based on Phonetic Features" *Proc. of the IEEE Dual-Use Technologies and Applications Conference*, May 1995, SUNY Inst. of Tech., Utica/Rome, pp. 310-315.
- [4] Bitar, N. and Espy-Wilson, C. Y. "Speech Parameterization Based on Phonetic Features: application to speech recognition." Eurospeech. September 1995, Madrid, Spain, pp. 1411-1414.
- [5] Bitar, N. N. and Espy-Wilson, C. Y. "Knowledge-Based Parameters for HMM Speech Recognition." ICASSP 1996, pp. 29-32.
- [6] Bitar, Nabil N. *Acoustic Analysis and Modeling of Speech Based on Phonetic Features*. Ph. D. thesis, Boston University, 1997.
- [7] Bitar, N. N. and Espy-Wilson, C. Y. "The Design of Acoustic Parameters for Speaker-Independent Speech Recognition" Eurospeech, September 1997, Rhodes, Greece, pp. 1239-1242

- [8] Carlson, Rolf, and Glass, James. "Vowel Classification Based on Analysis-by-Synthesis." Paper Th.sAM.3.1, ICSLP 92, Banff, Canada, pp. 575-578.
- [9] Chari, Venkatesh. *Extraction of Formant Frequencies by Adaptive Enhancement of Fourier Spectra*. MS thesis, EE, Boston University, 1992.
- [10] Choi, Jeung-Yoon, et al. "Labeling a speech database with landmarks and features." JASA 102():3163, December 1997.
- [11] Choi, Jeung-Yoon. *Lexical Access Project Tutorial: Labeling with Features*. unpublished, 2 February 1999.
- [12] Choi, Jeung-Yoon. *Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System*. Ph. D. thesis, EECS, MIT, June 1999.
- [13] Chun, Raymond Y. T. *A Hierarchical Feature Representation for Phonetic Classification*. M. Eng. thesis, EECS, MIT, March 1996.
- [14] Cole, R. A., Noel, M., Lander, T., and Durham, T. "New telephone speech corpora at CSLU." Eurospeech, vol 1, pp. 821-824, September 1995.
- [15] Crystal, T. H. and House, A. S. "Articulation rate and the duration of syllables and stress groups in connected speech." JASA 88(1):101-112, July 1990.
- [16] De Mori, Renato, and Biordano, Giovanna. "Algorithms for syllabic hypothesization in continuous speech." *Pattern Recognition* 14(1-6):245-260, 1981.
- [17] De Mori, Renato. *Computer models of speech using fuzzy algorithms*. New York: Plenum Press, 1983.
- [18] Di Benedetto, Maria-Gabriella. "Vowel Representation: Some observations on temporal and spectral properties of the first formant frequency." JASA 86(1):55-66, July 1989.
- [19] Duda, Richard O. and Hart, Peter E. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

- [20] formant(1-esps) manual pages. Entropic Laboratories, 1988-1990.
- [21] Erman, Lee D. and Lesser, Victor R. "The Hearsay-II Speech Understanding System: A Tutorial." In [48], pp. 361-381.
- [22] Espy-Wilson, Carol Y. *An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels*. Ph. D. thesis, Massachusetts Institute of Technology, 1987.
- [23] Espy-Wilson, Carol Y. "Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English." *JASA* 92(2): 736-757, August 1992.
- [24] Fakotakis, N., Tsopanoglou, A., and Kokkinakis, G. "A text-independent speaker recognition system based on vowel spotting." *Speech Communication* 12(1):57-68, 1993.
- [25] Fant, Gunnar. *Acoustic Theory of Speech Production*. Mouton & Company, 1960.
- [26] Fisher, W. Program TSYLB (version 2 revision 1.1), NIST. 7 August 1996.
- [27] Gillick, L. and Cox, S. J. "Some statistical issues in the comparison of speech recognition algorithms." *ICASSP 89*, paper S10b.5, pp. 532-535.
- [28] Godfrey, McDaniel and Holliman. "SWITCHBOARD: A Telephone Speech Corpus for Research and Development." 1992 *ICASSP Proceedings*.
- [29] Goldberg, H., Reddy, R., and Gill, G. "The ZAPDASH parameters, feature extraction, segmentation, and labeling for speech understanding systems." In *Summary of the CMU Five-year ARPA effort in speech understanding research*, Technical Report of the CMU Computer Science Speech Group, Carnegie-Mellon University, 1977.
- [30] Gow, David W., Melvold, Janis, and Manuel, Sharon. "How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing." *ICSLP 96*, vol. 1, pp. 66-69.

- [31] Green, P. D., Kew, N. R., and Miller, D. A. "Speech representations in the SYLK recognition project." in *Visual Representations of Speech Signals*, Cook, M., Beet, S., and Crawford, M., eds. Wiley, 1993.
- [32] Greenberg, S. "From sound to meaning: A syllable-centric perspective on spoken language" Third International Conference on Cognitive and Neural Systems, 28 May 1999.
- [33] Greenberg, S. and Kingsbury, B. "The modulation spectrogram: in pursuit of an invariant representation of speech." ICASSP 97 pp. 1647-1650.
- [34] Greenberg, S. "The Switchboard Transcription Project" in Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing. Johns Hopkins University.
- [35] Halberstadt, Andrew K. *Heterogeneous acoustic measurements and multiple classifiers for speech recognition*. Ph. D. thesis, EECS. Massachusetts Institute of Technology, 1999.
- [36] Hermansky, H., and Morgan, N. "RASTA processing of speech." IEEE Trans. SAP 2(4):578-589, October 1994.
- [37] Hermes, D. J. "Vowel onset detection." JASA 87(2):866-873. February 1990.
- [38] Huang, Caroline B. *An Acoustic and Perceptual Study of Vowel Formant Trajectories in American English*. RLE Technical Report 563, Research Laboratory of Electronics, MIT, March 1991.
- [39] Huffman, Marie K. and Krakow, Rena A. *Nasals, nasalization and the velum*. San Diego: Academic Press, 1993.
- [40] Hunt, M. J., Lennig, M., and Mermelstein, P. "Experiments in syllable-based recognition of continuous speech." ICASSP 1980, vol. 3, pp. 880-883.

- [41] Kasuya, H. and Wakita, H. "An approach to segmenting speech into vowel- and nonvowel-like intervals." *IEEE Trans. ASSP* 27(4):319-327, August 1979.
- [42] Kahn, D. "Syllable-based Generalizations in English Phonology." Ph. D. thesis, MIT, 1976.
- [43] Keyser, S. J. and Stevens, K. N. "Feature geometry and the vocal tract." *Phonology* 11, pp. 207-236.
- [44] Klatt, D. H. "Review of the ARPA Speech Understanding Project." *JASA* 62(6):1345-1366, December 1977.
- [45] Klatt, Dennis H. "Overview of the ARPA Speech Understanding Project" in [48], pp. 249-271.
- [46] Klatt, D. H. "Representation of the first formant in speech recognition and in models of the auditory periphery." *Proceedings Montreal Satellite Symposium on Speech Recognition*, 12th International Congress on Acoustics, Toronto, July 1986.
- [47] Lamel, L. et al. "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus." *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, 100-109.
- [48] Lea, W. A., ed. *Trends in Speech Recognition*, Prentice-Hall, 1980.
- [49] Lea, W. A. "Speech Recognition: Past, Present, and Future." in [48], pp. 39-98.
- [50] Lea, W. A. and Shoup, June E. "Specific Contributions of the ARPA SUR Project." in [48], pp. 382-421.
- [51] Lea, W. A. "Speech Recognition: What is Needed Now?" in [48], pp. 562-569.
- [52] Leung, Hong C. *The use of artificial neural networks for phonetic recognition*. Ph. D. thesis, EECS, Massachusetts Institute of Technology, 1989.

- [53] Lippmann, R. P. "Speech recognition by machines and humans." *Speech Communication* 22(1):1-15, 1997.
- [54] Liu, Sharlene A. "Landmark detection for distinctive feature-based speech recognition." *JASA* 100(5):3417-3430, November 1996.
- [55] Lowerre, Bruce and Reddy, Raj. "The Harpy Speech Understanding System." in [48], pp. 340-360.
- [56] Marshall, C. W. and Nye, P. W. "Stress and vowel duration effects on syllable recognition." *JASA* 74(2):433-443, August 1983.
- [57] Marslen-Wilson, William D. "Functional parallelism in spoken word recognition." in *Spoken Word Recognition*, ed. Frauenfelder and Tyler. Cambridge: MIT Press, 1987.
- [58] Medress, M. F. et al. "An automatic word spotting system for conversational speech." *ICASSP* 78, pp. 712-717.
- [59] Meng, Helen. *The Use of Distinctive Features for Automatic Speech Recognition*. MS thesis. EECS, MIT, September 1991.
- [60] Meng, Helen M. and Zue, Victor W. "Signal representation comparison for phonetic classification." Paper S5.9, *ICASSP* 91, pp. 285-288, Toronto, Canada, May 1991.
- [61] Mermelstein, P. "Automatic segmentation of speech into syllabic units." *JASA* 58(4):880-883, October 1975.
- [62] Mermelstein, P. "Recognition of monosyllabic words in continuous sentences using composite word templates." *ICASSP* 78, pp. 708-711.
- [63] Mermelstein, P., private communication.
- [64] Moll, Kenneth L. "Velopharyngeal Closure on Vowels." *Journal of Speech and Hearing Research* 5(1):30-37, March 1962.

- [65] Patel, A. D. *A biological study of the relationship between speech and music*. Ph. D. thesis, Harvard University, May 1996.
- [66] Peterson, G. E. and Barney, H. L. "Control Methods Used in a Study of Vowels." *JASA* 24(2):175-184, March 1952.
- [67] Pierce, J. R. "Whither Speech Recognition?" *JASA* 46(4):1049-1051, October 1969.
- [68] Pitrelli, John F. *Hierarchical Modeling of Phoneme Duration: Application to Speech Recognition* Ph. D. thesis, EECS, MIT, May 1990.
- [69] Pfitzinger, H. R., Burger, S., and Heid, S. "Syllable detection in read and spontaneous speech." *ICSLP 96*, vol. 2, pp. 1261-1264.
- [70] Rabiner, L. R. "On the application of energy contours to the recognition of connected word sequences." *AT&T Bell Laboratories Technical Journal* 63(9):1981-1995, November 1984.
- [71] Rabiner, L. R., Juang, B.-H., and Lee, C.-H. "An Overview of Automatic Speech Recognition." in *Automatic Speech & Speaker Recognition: Advanced Topics* Kluwer, 1996, pp. 1-30.
- [72] Reichl, W. and Ruske, G. "Syllable segmentation of continuous speech with artificial neural networks." *EuroSpeech 93*, vol. 3, pp. 1771-1774, September 1993.
- [73] Salomon, Ariel and Espy-Wilson, Carol. "Automatic detection of manner events based on temporal parameters." Paper S13.P02.9 at *Eurospeech 99*, Budapest, Hungary, September 1999.
- [74] Secrest, B. G. and Doddington, G. R. "An integrated pitch tracking algorithm for speech systems." *Proceedings ICASSP 1983*, pp. 1352-1355.
- [75] Seneff, S. "A computation model for the peripheral auditory system: application to speech recognition research." *ICASSP 1986*, Tokyo.

- [76] Shastri, L., Change, S., and Greenberg, S. "Syllable detection and segmentation using temporal flow neural networks." ICPHS-99, San Francisco, August 1999.
- [77] Shire, Michael Lee *Syllable Onset Detection from Acoustics*. MS Thesis, U. California at Berkeley, May 1997.
- [78] Smith, A. Richard and Sambur, Marvin R. "Hypothesizing and Verifying Words for Speech Recognition." in [48], pp. 139-165.
- [79] Stern, R. M., Acero, A., Liu, F. H., and Oshima, Y. "Signal processing for robust speech recognition." in *Automatic Speech & Speaker Recognition: Advanced Topics* Kluwer, 1996, pp. 357-384.
- [80] Stevens, K. N. "Models of Phonetic Recognition II: An Approach to Feature-based Recognition." Paper 5.5, Montreal Symposium on Speech Recognition. Proceedings. McGill University, 1986.
- [81] Stevens, K. N. "Phonetic Features and Lexical Access." Paper 10, Second Symposium on Advanced Man-Machine Interface Through Spoken Language. Makaha, Hawaii, 1988.
- [82] Stevens, K. N. "Phonetic evidence for hierarchies of features." *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pp. 242-258. Cambridge, Cambridge University Press, 1994.
- [83] Stevens, K. N. *Acoustic Phonetics*. Cambridge: MIT Press, 1998.
- [84] Stevens, K. N. Draft of lexical access paper. unpublished, 23 June 1999.
- [85] Strom, Nikko. *The NICO Artificial Neural Network Toolkit*.
<http://www.speech.kth.se/NICO/index.html>
- [86] Sun, Walter. "Analysis and interpretation of glide characteristics in pursuit of an algorithm for recognition." Masters thesis, MIT, November 1996.

- [87] Waibel, Alex. *Prosody and Speech Recognition*. London: Morgan Kaufmann Publishers, 1988.
- [88] Weinstein, C. J., McCandless, S., Mondschein, L. F., and Zue, V. W. "A system for acoustic-phonetic analysis of continuous speech." *IEEE Trans. ASSP* 23(1):54-67, February 1975.
- [89] Wolf, Jared J. and Woods, William A. "The HWIM Speech Understanding System." in [48], pp. 316-339.
- [90] Wu, S-L., Shire, M. L., Greenberg, S., and Morgan, N. "Integrating syllable boundary information into speech recognition." *ICASSP 97*, vol 2, pp. 987-990.
- [91] Wu, S-L., Kingsbury, B., Morgan, N., and Greenberg, S. "Incorporating information from syllable-length time scales into automatic speech recognition." *ICASSP 98*, vol 2, pp. 721-724.
- [92] Zhang, Yong. *Toward Implementation of a Feature-Based Lexical Access System*. M. Eng. thesis, MIT, March 1988.
- [93] Zwicker, E., Terhardt, E., and Paulus, E. "Automatic speech recognition using psychoacoustic models." *JASA* 65(2):487-498, February 1979.