Order Number 9509198

Invariant speech recognition and auditory object formation: Neural models and psychophysics

Govindarajan, Krishna K., Ph.D.

Boston University, 1995

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

BOSTON UNIVERSITY

GRADUATE SCHOOL

Dissertation

INVARIANT SPEECH RECOGNITION AND AUDITORY OBJECT FORMATION: NEURAL MODELS AND PSYCHOPHYSICS

by

KRISHNA K. GOVINDARAJAN

B.S., University of Colorado, 1989

Submitted in partial fulfillment of the

requirements for the degree of

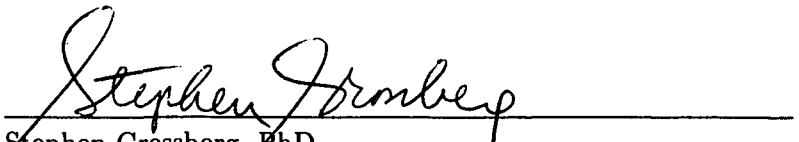
Doctor of Philosophy

1995

Approved by

First Reader 

Michael A. Cohen, PhD
Associate Professor of Cognitive and Neural Systems
and Computer Science

Second Reader 

Stephen Grossberg, PhD
Wang Professor of Cognitive and Neural Systems

Third Reader 

Gail A. Carpenter, PhD
Professor of Cognitive and Neural Systems

Acknowledgments

Life contains a mixture of randomness and regularities, and science's goal is to find the regularities in life. I would like to thank those individuals that have shown me the patterns, those that have shown me randomness, and to those rare individuals who gave me both.

I would like to thank my readers, Prof. Michael Cohen, who collaborated on the research in Chpt. 3 and 4, Prof. Stephen Grossberg, who collaborated on the research in Chpt. 4, and Prof. Gail Carpenter, who collaborated on the research in Chpt. 2, for their invaluable guidance, knowledge, and intuition in performing this research. In addition, I would like to thank Lonce Wyse for his seminal contributions to the work in Chpt. 4. I would also like to thank other professors, Dan Bullock, Laurel Carney, Carol Espy-Wilson, and Ennio Mingolla, who provided insights and provoked, such as "What are the units ?"

I would like to thank the following people for advise and moral assistance during my dissertation, my housemates, the Barnies: Andy Worth, Dan Cruthirds, Alan Gove, Doug Greve, Rod Rinkus, Gary Bradski, David Rosen, Mukund Balasubramaniam, John Reynolds, and Chad Lewis; my friends, coworkers and officemates at the CNS department, especially Frank Guenther, Chris Ra, Lonce Wyse, Steve Lehar, Ian Boardman, Sulochana Naidoo, Sonya Rost, Alison Dorsky, Niall McLoughlin, Luiz Pessoa, Paulo Gaudiano, Bruce Fischl, Dave Johnson, Josh Krieger, Steve Olson, Karen Roberts, Ah-Hwee Tan, and Bill Woods.

Most importantly, I would like to express my undying gratitude to my parents, my brother Prasad, Sujatha, Bhaskar, my grandparents S. V. Thatha, Calcutta Thatha, Ceelama, and Srirangama, for their emotional support, and for providing the inspiration to pursue my PhD.

This research was supported in part by the Air Force Office of Scientific Research

(AFOSR-F49620-92-J-0225), DARPA (AFOSR-90-0083), the National Science Foundation (NSF- IRI-90-00530), a Graduate Research Assistant Scholarship (GRASP) and for one semester by a Teaching Fellowship.

INVARIANT SPEECH RECOGNITION AND AUDITORY OBJECT FORMATION: NEURAL MODELS AND PSYCHOPHYSICS

(Order No.)

KRISHNA K. GOVINDARAJAN

Boston University Graduate School, 1995

Major Professor: Michael A. Cohen, PhD,
Associate Professor of Cognitive and Neural Systems

Abstract

This dissertation investigates three topics concerning variability and robustness in speech perception: variability of the speech signal across speakers, variability due to speaking rate effects, and the robustness of speech perception in noisy environments.

Given that the speech signal corresponding to a given phoneme can vary considerably across speakers, invariant speech perception can be facilitated by normalizing the signal across speakers. In chapter 1, 160 intrinsic and extrinsic speaker normalization methods are compared using a neural network, fuzzy ARTMAP, and K-Nearest Neighbor (K-NN) categorizers trained and tested on disjoint sets of speakers of the Peterson-Barney vowel database. ARTMAP and K-NN show similar trends, with K-NN performing better but requiring about ten times as much memory. The optimal intrinsic normalization method is bark scale using the differences between all frequencies, while the optimal extrinsic method is linear transformation of the vowel space to a canonical representation.

In chapter 2, psychophysical studies of adaptation to the mean silence duration between two different stop consonants are examined. Using natural speech stimuli, the first experiment shows that the category boundary between hearing only one

or hearing both stop consonants varied as a function of the distribution of silent intervals. The second experiment shows that the variance of the distribution did not significantly affect the boundary, and the final experiment shows sequential effects in the adaptation process. Finally, a model of the adaptation process is developed which emulates the data.

In environments with multiple sound sources, the auditory system is capable of teasing apart the impinging jumbled signal into different mental objects. Chapter 3 presents a neural network model of auditory scene analysis, which groups different frequency components based on pitch and spatial location cues and allocates the components to different objects. While location primes the grouping mechanism, segregation is based solely on harmonicity. The model qualitatively emulates results from psychophysical grouping experiments, such as how a tone sweeping upwards in frequency groups due to frequency proximity with a downward sweeping tone even if noise exists at the intersection point; and illusory percepts, such as the illusion of a tone continuing through noise.

Contents

1 Introduction 1

2 Speaker normalization methods for vowel recognition: Comparative analysis using neural network and nearest neighbor classifiers 4

2.1 Introduction 4

2.2 Speaker normalization 8

2.3 Normalization methods 9

2.3.1 Intrinsic normalization methods 9

2.3.2 Extrinsic normalization methods 11

2.4 Peterson-Barney vowel database 14

2.5 Algorithms for comparing normalization methods 15

2.5.1 Fuzzy ARTMAP 15

2.5.2 K-Nearest Neighbor 17

2.6 Comparative evaluation of speaker normalization methods 18

2.6.1 Intrinsic methods 18

2.6.2 Extrinsic methods 18

2.7 Discussion 27

2.7.1 Fuzzy ARTMAP vs. K-NN 27

2.7.2 Differences between vowel space scales 32

2.7.3 Intrinsic methods 32

2.7.4 Extrinsic methods 33

2.8 Summary 33

3 A psychophysical study of adaptation to silent intervals during variable-rate speech **35**

3.1 Introduction 35

3.1.1 Cue adaptation 36

3.1.2 Repp (1980) experiment 37

3.2 Experiment 1: Replication of Repp (1980) cluster condition 41

3.2.1 Subjects 41

3.2.2 Stimuli 41

3.2.3 Procedure 43

3.2.4 Results and discussion 43

3.3 Experiment 2: Different place of articulation 47

3.3.1 Subjects 47

3.3.2 Stimuli 47

3.3.3 Procedure 47

3.3.4 Results and discussion 47

3.4 Experiment 3: Effects of silent interval distribution variance 51

3.4.1 Subjects 51

3.4.2 Stimuli 51

3.4.3 Procedure 52

3.4.4 Results and discussion 52

3.5 Experiment 4: Temporal characteristics of adaptation 56

3.5.1 Subjects 56

3.5.2 Stimuli 56

3.5.3 Procedure 57

3.5.4 Results and discussion 57

3.6 Linear adaptation model 58

3.6.1 Window length and weighting 62
3.7 General results and conclusion 65

4 A neural network model of auditory scene analysis 69
4.1 Introduction 69
4.1.1 Auditory scene analysis 69
4.1.2 Grouping principles 71
4.1.3 Primitive versus schema-based segregation 73
4.2 Grouping cues 73
4.2.1 Temporal and frequency separation 74
4.2.2 Continuity illusion 75
4.2.3 Harmonicity and pitch 75
4.2.4 Bounce and cross percept in crossing glide complexes 78
4.2.5 Spatial location 80
4.2.6 Amplitude modulation (AM) 84
4.2.7 Frequency modulation (FM) 85
4.2.8 Onsets and offsets 85
4.3 Existing models of segregation 87
4.4 Model of auditory streaming and grouping 88
4.4.1 Auditory peripheral processing 90
4.4.2 Spectral stream layer 91
4.4.3 Pitch summation layer 96
4.4.4 Pitch stream layer 96
4.5 Simulation results of model 98
4.5.1 Inharmonic simple tones 98
4.5.2 Continuity illusion 103
4.5.3 Bounce percepts for crossing glides 111

4.5.4 Steiger (1980) diamond stimulus 118
4.6 Extension to model: Spatial location 123
4.6.1 Spatial location cues 123
4.6.2 Extended model 125
4.7 Summary 129

References **131**

List of Tables

2.1 Correct vowel identification rates for mixed and blocked speakers. After Nearey (1989). 9

2.2 Vowels used in the Peterson and Barney (1952) study. 14

2.3 Fuzzy ARTMAP test set performance with intrinsic normalization. Numbers in parentheses give the average number of F_2^a nodes after training. Vowel space scales: N = nonscaled, B = bark scaled, Be = bark scaled with end-correction, M = mel scaled, ERB = equivalent rectangular bandwidth scaled, log_e = natural logarithm scaled, $log_{1.06}$ = semitone scaled, and log_{10} = log base 10 scaled. Intrinsic normalization methods 1-8 use only the first two formants $[F_1', F_2']$; methods 9-16 use $[F_0', F_1', F_2', F_3']$; for methods 17-32, differences between the transformed $F_0', \ldots, F_3'$ were computed. Methods 17-24 (Diff Subset) employ the three differences $F_1' - F_0', F_2' - F_1', F_3' - F_2'$; methods 25-32 (Diff All) employ all six differences $F_1' - F_0', F_2' - F_0', F_3' - F_0', F_2' - F_1', F_3' - F_1', F_3' - F_2'$. Fuzzy ARTMAP simulation parameters: $\bar{\rho}_a = 0.0$, $\alpha = 0.1$, $\beta = 1.0$. 19

2.4 L_1 and L_2 K-NN (K = 10) test set performance with intrinsic normalization. The vowel space scales are specified in Table 2.3. 20

2.5 Fuzzy ARTMAP test set performance with centroid subtraction across all frequencies (CS) extrinsic normalization. 22

2.6 L_1 and L_2 K-NN test set performance with centroid subtraction across all frequencies (CS) extrinsic normalization. 23

2.7 Fuzzy ARTMAP test set performance with centroid subtraction for each frequency (CSi) extrinsic normalization. 23
2.8 L_1 and L_2 K-NN test set performance with centroid subtraction for each frequency (CSi) extrinsic normalization. 25
2.9 Fuzzy ARTMAP test set performance with linear scale (LS) extrinsic normalization. 27
2.10 L_1 and L_2 K-NN test set performance with linear scale (LS) extrinsic normalization. 28
2.11 Fuzzy ARTMAP test set performance with linear transformation (LT) extrinsic normalization. 28
2.12 L_1 and L_2 K-NN test set performance with linear transformation (LT) extrinsic normalization. 30

List of Figures

2·1 Vowel space (mean F_1 vs. F_2) of all 76 speakers, and of 33 males, 28 females, and 15 children, for the ten vowels of the Peterson and Barney (1952) database. 5

2·2 Fuzzy ARTMAP architecture with its three component modules: ART_a, ART_b, and the map field F^{ab}. The ART_a module transforms the M-dimensional input vector **a** into a 2M-dimensional input vector $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c)$ at the F_0^a field through complement coding. **A** is the input to the ART_a field F_1^a. Similarly, the input to the ART_b field F_1^b is the vector $\mathbf{B} = (\mathbf{b}, \mathbf{b}^c)$. The activation of field F_1^a causes a category to become active at F_2^a, which in turn leads to a prediction by ART_a at the map field F^{ab}. If the prediction is disconfirmed at ART_b, inhibition of the map field activity starts the match tracking process: the ART_a vigilance ρ_a is raised slightly over the match ratio. This in turn causes an ART_a search which leads to either the activation of another existing ART_a category that correctly predicts **B**, or a new uncommitted ART_a category node. 16

2·3 Comparison between fuzzy ARTMAP and K-NN for intrinsic normalization methods. Normalization identification numbers are in Tables 2.3 and 2.4. Methods 26 and 27 (B and Be, Diff All) have the best performance. 21

2·4 Comparison between fuzzy ARTMAP and K-NN for the centroid subtraction across all frequencies (CS) extrinsic normalization method (Tables 2.5 and 2.6). Fuzzy ARTMAP performed best with methods 26 and 27 (B and Be, Diff All), while K-NN performed best with method 13 (ERB, $[F_0'', F_1'', F_2'', F_3'']$). 24

2·5 Comparison between fuzzy ARTMAP and K-NN for the centroid subtraction for each frequency (CSi) extrinsic normalization method (Tables 2.7 and 2.8). Fuzzy ARTMAP performed best with method 26 (B, Diff All), and K-NN performed best with methods 14 through 16 ($log_e/log_{1.06}/log_{10}$, $[F_0'', F_1'', F_2'', F_3'']$). 26

2·6 Comparison between fuzzy ARTMAP and K-NN for the linear scale (LS) extrinsic normalization method (Table 2.9 and 2.10). Fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN performed best with methods 26 and 27 (B and Be, Diff All), method 29 (ERB, Diff All), and method 18 (B, Diff Subset), respectively. 29

2·7 Comparison between fuzzy ARTMAP and K-NN for the linear transformation (LT) extrinsic normalization method (Tables 2.11 and 2.12). Fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN performed best with methods 26 (B, Diff All), method 12 (M, $[F_0'', F_1'', F_2'', F_3'']$), and method 13 (ERB, $[F_0'', F_1'', F_2'', F_3'']$), respectively. 31

3·1 Spectrogram of intervocalic stop consonant /ada/ uttered by a male speaker. 36

3·2 Schematic representation of the /ib-ga/ cluster token from Repp (1980). 37

3·3 Distribution of the silent intervals for the three anchor conditions for the cluster case of Repp (1980). 39

3·4 Results from the Repp (1980) experiment for the three anchor conditions for both the cluster and geminate case as averaged over 8 subjects. The datapoints were estimated from Figure 2 of Repp (1980). 40
3·5 Results for experiment 1 for the six subjects for the low, no, and high anchor conditions. The figures show the average number of stop consonants heard for each silent interval. 45
3·6 Results from experiment 1 pooled across the subjects for the low, no, and high conditions. 46
3·7 Results for experiment 2 for the six subjects for the three anchor conditions. The figures show the average number of stop consonants heard for each silent interval. 49
3·8 Results from experiment 2 pooled across the subjects for the three conditions. 50
3·9 Distribution of the silent intervals for the small range condition, the U anchor condition, and the normal range condition. 53
3·10 Results for experiment 3 for the six subjects for the small range, U anchor and the normal range. The figures show the average number of stop consonants for each silence duration. 54
3·11 Results from experiment 3 pooled across the subjects for the three conditions. 55
3·12 Results for experiment 4 for the five subjects for the different conditions. The curves correspond to the different P' conditions, as listed. 59
3·13 Results from experiment 4 pooled across the subjects for the different conditions. The curves correspond to the different P' conditions, as listed. 60

3·14 Errors as a function of window size for each subject's data from experiment 2. 63
3·15 Errors as a function of window size for each subject's data from experiment 3. 64
3·16 Weighting coefficients a_k for a seven token window length (M = 7) for each subject's data from experiment 2. 66
3·17 Weighting coefficients a_k for a seven token window length (M = 7) for each subject's data from experiment 3. 67

4·1 A groups better with B if they are closer in frequency. However, simultaneous cues, such as common onsets, common offsets and harmonicity, can help group B and C. After Bregman and Pinker (1978). 70
4·2 Stimulus and percept of the continuity illusion. (a) shows the stimulus that is presented to listeners, and (b) represents the percept. Note that in the stimulus, the tone does not continue through the noise, but stops at the onset of the noise, and continues at the offset of the noise, but the percept is that the tone continues through the noise. . 72
4·3 When A and B are presented by themselves, listeners could easily judge the order of them. If A and B were flanked by tones F, then listeners had a more difficult time. However, if the captor tones C surrounded the flankers, then F streamed with C, leaving A-B to a different stream, allowing the listeners to hear the order once again. After Bregman and Rudnicky (1975). 74

4·4 Stimuli and percept of the experiment by Steiger (1980). (a) and (b) show the stimuli that were presented to the subjects. In (b), the noise spectra is not added to the glides, but actually replaces the glide portions. For both the stimuli in (a) and (b), listeners hear the two streams shown in (c) and (d). In (b), a third stream is heard corresponding to the broadband noise bursts. After Steiger (1980). . . 76

4·5 Listeners are presented with a cyclic pattern of high (H) and low (L) tones, which are either connected (a), or point towards each other (b), or have no trajectory between them (c). The effect is that listeners heard one stream in (a) and two streams in (c), but that there was a higher probability of hearing one stream in (b), where they are pointing towards each other. After Bregman and Dannenbring (1973). 77

4·6 Stimuli and listeners' responses in Halpern (1977) for different harmonic conditions. The complex glides were all 1 second long, and the numbers next to a glide is its harmonic number. The numbers below each figure corresponds to the preference of hearing a bounce or a cross: numbers greater than 2.5 correspond to a bounce percept, and numbers below 2.5 correspond to a cross percept. After Halpern (1977). 79

4·7 Stimuli of Tougas and Bregman (1990) for four different harmonic conditions. All but the rich crossing condition produced a bounce percept, even when the interval I was filled with silence, noise, or just the glides. The order, from greatest to the least, of bounciness was rich bouncing, all pure, and all rich. After Tougas and Bregman (1990). 81

4·8 (a) Scale illusion in which a downward and an upward scale are being played at the same time, except that every other tone in a given scale is presented to the opposite ear, corresponding to an L or R for left and right ear. (b) The result is that listeners group based on frequency proximity, and heard the two streams A and B. After Deutsch (1975). 83

4·9 (a) If a harmonic at 500 Hz started 240 ms before the rest of a short synthetic vowel, then it has a diminished contribution to the vowel identity. (b) If a 1000 Hz tone was added that started at the same time as the 500 Hz harmonic and stopped at the vowel, the harmonic's contribution increases slightly due to the grouping of the 500 and 1000 Hz tones. After Darwin and Sutherland (1984). 87

4·10 Block diagram of the auditory streaming model. 89

4·11 Interaction between the energy measure, the spectral stream layer, the pitch stream layer, and the pitch summation layer. The energy measure layer is fed forward in a frequency-specific one-to-many manner to each frequency-specific stream node in the spectral stream layer. In addition, this feed-forward activation is contrast-enhanced. There is also competition within the spectral stream layer across streams for each frequency so that a component is allocated to only one stream at a time. Each stream in the spectral stream layer activates its corresponding pitch stream in the pitch stream layer. Each pitch neuron receives excitation from its harmonics in the corresponding stream. Since each pitch stream is a winner-take-all network, only one pitch can be active at any given time. Across streams in the pitch stream layer, there is asymmetric competition for each pitch so that one stream is biased to win and the same pitch can not be represented in another stream. Finally, the winning pitch neuron feeds back excitation to its harmonics in the corresponding spectral stream. The stream also receives non-specific inhibition from the pitch summation layer, which sums up the activity at the pitch stream layer for that stream. This non-specific inhibition helps to suppress those components that are not supported by the top-down excitation, which plays the role of a priming stimulus or expectation (Carpenter & Grossberg, 1991). 92

4·12 Stimuli and the listeners' percepts that the model is capable of emulating. The hashed boxes represent broadband noise. The stimuli consist of: (a) two inharmonic tones, (b) tone-silence-tone, (c) tone-noise-tone, (d) a ramp or glide-noise-glide, (e) crossing glides, (f) crossing glides where the intersection point has been replaced by silence; (g) crossing glides where the intersection point has been replaced by noise, (h) Steiger (1980) diamond stimulus, and (i) Steiger (1980) diamond stimulus where bifurcation points have been replaced by noise. 99

4·13 (a) spectrogram and (b) result of energy measure for the two tone stimulus. 100

4·14 Model results for the two tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 101

4·15 Schematic of how the model segregates the two inharmonic tones into two different streams. See text for explanation. 102

4·16 (a) spectrogram and (b) result of energy measure for the tone-silence-tone stimulus. 104

4·17 Model results for the tone-silence-tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 105

4·18 (a) spectrogram and (b) result of energy measure for the tone-noise-tone stimulus. 106

4·19 Model results for the tone-noise-tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 107

4·20 The (a) spectral and (b) pitch stream layers for stream 3 for the tone-noise-tone stimulus. 108
4·21 (a) spectrogram and (b) result of energy measure for the ramp stimulus. 109
4·22 Model results for the ramp stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 110
4·23 (a) spectrogram and (b) result of energy measure for the crossing glide stimulus. 112
4·24 Model results for the crossing glide stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 113
4·25 (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with silence replacing the intersection point. 114
4·26 Model results for the crossing glide stimulus with silence replacing the intersection point. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 115
4·27 (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with noise replacing the intersection point. 116
4·28 Model results for the crossing glide stimulus with noise replacing the intersection point. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 117
4·29 (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus. 119

4·30 Model results for the Steiger (1980) diamond stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 120
4·31 (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points. . 121
4·32 Model results for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2. 122
4·33 Geometric representation of spatial lateralization using interaural timing differences (ITD). 124
4·34 Block diagram of the extended streaming model. 125
4·35 Interaction between spatial locations in the f-τ field, pitch stream layer, and the spectral stream layer. The non-specific inhibitory neurons are not shown. Only one stream can occupy one spatial location, except at the central "head-centered" location $\tau = 0$, where multiple streams can be represented. Once a spatial location has been derived for all the components, the spatial location non-specifically primes all the neurons in its corresponding pitch stream layer. At the central location, the N streams are all primed. Once components have been grouped based on pitch, the neurons in a spectral stream layer specifically excite the components at their corresponding spatial location. At the central location, the spectral neurons, corresponding to a given frequency, from all N streams excite the corresponding neuron at $\tau = 0.127$

Chapter 1

Introduction

While machine speech recognition and an understanding of human speech perception have advanced over the last several decades, the variabilities in the speech signal still impede a veridical model of human speech perception and an adequate speech recognition device (Klatt, 1992). The variabilities in the acoustic signal include:

- contextual variability, e.g. coarticulation, the acoustic cue for /d/ is different for /da/ versus /di/,
- inter-speaker variability, e.g. dialects, different vocal tract sizes cause the speech signal corresponding to the same phoneme to be quite variable across speakers,
- intra-speaker variability, e.g. emotional state of the speaker, different speaking rates,
- and variability due to environmental conditions, e.g. room reverberations, microphone, background noise, other speakers.

Three research problems addressing variability and robustness in speech perception are examined in this dissertation. The first research topic relates to inter-speaker variability, different vocal tract sizes; the second topic studies intra-speaker variability, speaking rate effects; and the third topic investigates variability due to environmental factors, background noise and other speakers.

The speech signal corresponding to a given phoneme can vary considerably across speakers, due to differential vocal tract sizes. To achieve invariant speech perception and recognition, the signal can be transformed to a more canonical representation, or normalized across speakers, for easier classification. In Chapter 2, 160 preprocessing, or normalization, methods are compared using a self-organizing neural network classifier, fuzzy ARTMAP, and a nearest neighbor classifier, to determine which combinations of preprocessing and classification system provide the most accurate recognition system. The 160 methods are obtained by factorially varying eight different frequency scales with four combinations of the frequency components, and five speaker adaptation schemes.

Durational factors, such as silent intervals, act as acoustic cues in the perception of phonemes and word boundaries. During variable-rate speech perception and recognition, the distribution of these durational cues can influence how speech sounds are categorized for purposes of recognition. Thus, in order for a listener to consistently categorize these speech sounds, the listener must adapt to the speaker's speech rate, or perform normalization across time, and create a canonical representation for invariant identification. In Chapter 3, psychophysical studies using silent intervals are performed to understand the specific nature of this adaptation process. The results suggest that listeners base their judgments of these cues based on the mean silent interval, where these intervals vary as a function of speaking rate. In deriving this mean, listeners adapt to the mean interval within some time window.

In environments where there are multiple sound sources, listeners are able to segregate the different signals arising from these sources even though there is only one merged signal impinging upon the ear. This auditory scene analysis also helps listeners to hear a particular speaker in noisy environments and in environments with other speakers, e.g. at cocktail parties. It is thus a crucial step in segregating

speech sounds for purposes of perception and recognition. Two factors that influence segregation, among others, are the pitch of a sound and its spatial location. A key issue concerns how overlapping combinations of spectral components can be separated into the pitches and locations of different acoustic sources. A neural network model of auditory scene analysis that suggest how pitch and spatial location are used for segregation is presented in Chapter 4.

Chapter 2

Speaker normalization methods for vowel recognition: Comparative analysis using neural network and nearest neighbor classifiers

2.1 Introduction

Human listeners are able to identify as a single phoneme a wide variety of speech signals produced by different speakers in different contexts. For example, the vowel /æ/ is recognized despite the fact that the average F_1 formant frequency is approximately 660 Hz for males and 1020 Hz for children (Figure 2·1) (Peterson & Barney, 1952). In order for humans to consistently categorize speech sounds from multiple speakers, they must compensate for the variability of the speech signal across speakers. *Speaker normalization* denotes the process whereby a listener compensates for individual characteristics of a speech signal in order to extract invariant features needed to identify the sound. The two main classes of normalization methods are *intrinsic* and *extrinsic* (Ainsworth, 1975; Nearey, 1989). Intrinsic normalization uses only the information present in each vowel token. Extrinsic normalization uses information from several vowel tokens of a given speaker.

Procedures are developed here that can be used to make systematic comparisons of the many speaker normalization schemes that have been proposed in recent decades. To evaluate a given normalization method, the 1520 vowel token vectors, consisting of the fundamental (F_0) and first three formants (F_1, F_2, F_3) of the Peter-

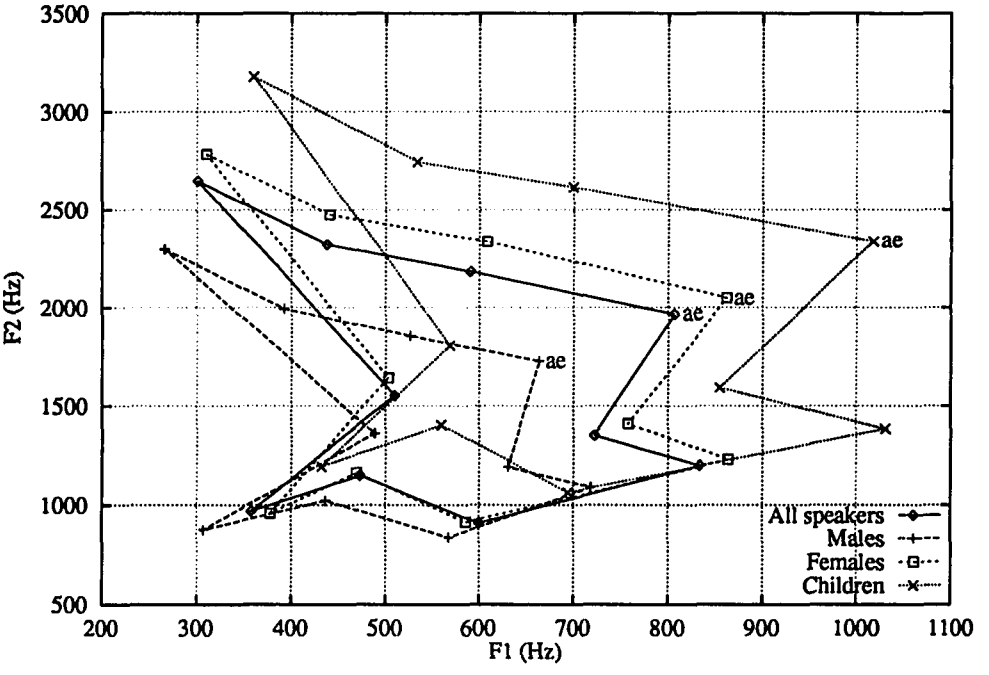


Figure 2·1: Vowel space (mean F_1 vs. F_2) of all 76 speakers, and of 33 males, 28 females, and 15 children, for the ten vowels of the Peterson and Barney (1952) database.

son and Barney (1952) database, are preprocessed using that method. Normalized inputs from about 30% of the speakers are used to train three different classifiers: a neural network, fuzzy ARTMAP (Carpenter, Grossberg, Markuzon, Reynolds, & Rosen, 1992) and two K-Nearest Neighbor (K-NN) systems (Dasarathy, 1991). The remaining test data set is then presented to each classifier, which tries to identify a test set input as one of ten vowel sounds. The normalization scheme in question is evaluated in terms of the number of correct test set identifications made by each of the classifiers. Speaker independence is required since the test set inputs and the training set inputs are generated by disjoint sets of speakers (men, women, and children). Comparative evaluations of 32 intrinsic and 128 extrinsic normalization schemes are carried out using this method.

For the intrinsic normalization schemes, eight scales were compared: one non-scaled scale; four psychophysical scales, bark scale, bark scale with end-correction, mel scale, and equivalent rectangular bandwidth ERB (scale); and three log measures, a semitone scale, natural log scale, and log base 10 scale. For each of these eight scales, four frequency combinations were tested by the categorizers: $[F_1', F_2']$ only; $[F_0', F_1', F_2', F_3']$; differences between all combinations of formants and F_0' (Diff All); and a subset of the differences between the formants and F_0' (Diff Subset).

For the extrinsic methods, speaker-specific adaptation was superimposed on each of the 32 intrinsic normalization methods. Four types of extrinsic normalization methods were tested: centroid subtraction across frequencies (CS), centroid subtraction for each frequency (CSi), linear scale (LS), and linear transformation (LT). The CS method subtracts the mean frequency ($\overline{F}$) across all frequencies from a speaker's set of vowels. The CSi method subtracts the mean frequency ($\overline{F_i}$) from its respective frequency in a speaker's set of vowels. The LS method computes the minimum and maximum value for each frequency across all the vowels of a given speaker, then

rescales every value for that frequency to the range [0,999]. Finally, the LT method adaptively computes a matrix for each speaker to warp the vowel space to the mean vowel space across all speakers.

The three pattern recognition systems (fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN) generally agreed on which normalization methods gave better predictive performance on test set data. K-NN tended to outperform fuzzy ARTMAP by a few percent, but also has greater storage requirements.

In general, the psychophysical measures outperformed the log measures. For all the intrinsic and extrinsic methods, fuzzy ARTMAP performed best using bark (or bark with end correction) Diff All. Although the K-NN categorizers' optimal performance varied more, the majority of the normalization methods performed best with the psychophysical measures. For the intrinsic methods, K-NN algorithms chose bark Diff All method. For the CS extrinsic method, they performed best with ERB $[F_0'', F_1'', F_2'', F_3'']$ method. For the LS method, L_1/L_2 K-NN chose ERB Diff All/bark Diff Subset. For the LT method, L_1/L_2 K-NN performed best with mel/ERB $[F_0'', F_1'', F_2'', F_3'']$. Finally, for the CSi method, the K-NN methods chose log $[F_0'', F_1'', F_2'', F_3'']$. Among the extrinsic methods, the order of performance consisted of the following, from best to worst: LT, CSi, LS, and CS.

A primary goal of this paper is to develop an efficient, standard method to compare and evaluate the many normalization methods in the literature. Other neural network approaches have compared a limited number of normalization methods, often without speaker independence in the training and test sets.

Section 2.2 reviews some data on vowel perception as it applies to speaker normalization. Section 2.3 discusses the intrinsic and extrinsic speaker normalization methods that were tested. Section 2.4 describes the Peterson-Barney database. Section 2.5 outlines the pattern recognition schemes, fuzzy ARTMAP and K-NN, used

to evaluate the normalization methods. Section 2.6 presents the response of the different categorizers to the different normalization methods, and Section 2.7 discusses these results.

2.2 Speaker normalization

A variety of psychophysical experiments illustrate how listeners employ speaker normalization. For example, Assmann, Nearey, and Hogan (1982) showed that listeners identify fixed duration steady-state vowels with 86.2% to 91.5% accuracy. Jenkins, Strange, and Edman (1983) showed that listeners perform at 88.2% when presented with variable duration steady-state vowels. Thus, listeners are able to accurately identify vowels without the benefit of transitional or durational information, even if formant frequencies for different vowels overlap.

A different type of evidence for speaker normalization derives from speaker adaptation data. In particular, the identity of a test vowel can be changed if the formants of vowels in the preceding carrier sentence are altered (Ladefoged & Broadbent, 1957; Ainsworth, 1975; Deschovitz, 1977; Nearey, 1978; Remez, Rubin, Nygaard, & Howell, 1987; Nearey, 1989). Ladefoged and Broadbent (1957) showed that the identity of a vowel in a test word (/bVt/) shifted when the formant frequency ranges of the preceding synthetic carrier sentence, "Please say what this word is ___," was modified. Remez et al. (1987) replicated Ladefoged and Broadbent's experiment using sinusoidal voices. Nearey (1978), using synthetic stimuli, showed that prior presentation of an adult or child /i/ shifted vowel categories along a continuum of F_1 and F_2 values: vowel boundaries shifted towards higher frequencies if the child /i/ preceded the formants. Deschovitz (1977), using natural speech stimuli, showed that listeners are more error prone when identifying /bVt/ words spoken by an adult male embedded within a child's carrier sentence, "Please say ___ for me." Similarly,

Stimuli	Mixed (%)	Blocked (%)	
/V/	57.4	68.8	Strange et al. (1976)
/pVp/	83.0	90.5	Strange et al. (1976)
/V/	92.2	98.5	Macchi (1980)
/tVt/	91.4	98.0	Macchi (1980)
/V/	94.6	95.9	Assmann et al. (1982)
Gated /V/	86.2	90.5	Assmann et al. (1982)

Table 2.1: Correct vowel identification rates for mixed and blocked speakers. After Nearey (1989).

Ainsworth (1975) found that synthetic carrier vowels / i u ɑ/ could influence the vowel categories' centers, with the vowel category boundaries shifting by as much as 16% depending on whether the listener perceived a male or a child speaker. Nearey (1989) evaluated the effects of F_0 and higher formants, as well as the range of F_1-F_2 frequencies on vowel perception. Nearey, using either a high or low frequency range for /i ɒV/, found that the higher formants had little influence while the ensemble range of frequencies had the most influence, followed by F_0, which had more influence on F_1 than on F_2.

Other adaptation experiments show fewer errors occur during blocked conditions, in which only one speaker's vowel tokens are presented within a trial, than during mixed conditions, in which the identity of the speaker varies randomly from token to token within the trial (Strange, Verbrugge, Shankweiller, & Edman, 1976; Macchi, 1980; Assmann et al., 1982; Nearey, 1989) (Table 2.1).

2.3 Normalization methods

2.3.1 Intrinsic normalization methods

For the intrinsic normalization schemes, eight normalization methods were compared: one nonscaled (N) scale; four psychophysical scales: bark scale (B) (Zwicker & Terhardt, 1980), bark scale with end-correction (Be) (Traunmüller, 1981), mel scale (M)

(O'Shaughnessy, 1987), and equivalent rectangular bandwidth scale (ERB) (Moore & Glasberg, 1983); and three log measures: a semitone scale ($log_{1.06}$), natural log scale (log_e), and log base 10 scale (log_{10}).

The bark scale is a psychophysically-derived measure, which is thought to correspond to internal bandpass filters, or critical bands. The critical band is the bandwidth of a filter in which acoustic energy is integrated. The bark scale (B) transforms $F_0 \ldots F_3$ to $F'_0 \ldots F'_3$ according to the equation:

$$F'_i = 13.0 * \arctan(0.76 * F_i/1000) + 3.5 * \arctan(F_i/7500)^2, \qquad (2.1)$$

where F_i is the i^{th} frequency, in Hz. Bark scale with end-correction (Be) adjusts the lower frequencies before converting to them to bark scale, frequencies below 150 Hz are increased to 150 Hz; frequencies between 150 and 200 Hz are reduced to $0.8F_i$+30; and frequencies between 200 and 250 Hz are increased to $1.2F_i - 50$. This factor was applied to correct for discrepancies at low frequencies between psychophysical data and equation 2.1.

The mel scale (M), which is a psychophysical scale that is derived based on the perceived pitch relationships of two tones, corresponds to the transformation:

$$F'_i = 2595 \log_{10}(1 + F_i/700). \qquad (2.2)$$

The fourth psychophysical scale is the equivalent rectangular bandwidth (ERB) scale. The ERB of a filter corresponds to the bandwidth of a rectangular filter, which passes the same amount of power. While the ERB scale is similar to the bark scale, the two scales differ for frequencies below 1000 Hz. The ERB scale is calculated by:

$$F'_i = 11.17 * \log_e((F_i + 312)/(F_i + 14675)) + 43. \qquad (2.3)$$

The three logarithmic measures consist of the semitone scale:

$$F_i' = \log_{1.06}(F_i), \tag{2.4}$$

the natural logarithm scale:

$$F_i' = \log_e(F_i), \tag{2.5}$$

and the log base 10 scale:

$$F_i' = \log_{10}(F_i). \tag{2.6}$$

Each of the eight normalization scales was tested with four different combinations: only the first two formants $[F_1', F_2']$; the fundamental and all three formants $[F_0', F_1', F_2', F_3']$; the three differences $F_1' - F_0', F_2' - F_1', F_3' - F_2'$ (Diff Subset); and all six difference combinations $F_1' - F_0', F_2' - F_0', F_3' - F_0', F_2' - F_1', F_3' - F_1', F_3' - F_2'$ (Diff All). Combining the 8 vowel space scales and the 4 frequency combinations, 32 intrinsic methods were tested.

Syrdal and Gopal (1986) , using the bark scale with end correction (Be) and the Diff Subset method, obtained a performance rate of 81.8% on the Peterson-Barney database, using linear discriminant analysis (LDA) with the U (jacknife) method.

2.3.2 Extrinsic normalization methods

For the extrinsic methods, adaptation to a speaker was superimposed on each of the 32 intrinsic normalization methods. Four types of extrinsic normalization were tested: centroid subtraction across frequencies (CS), centroid subtraction for each frequency (CSi), linear scale (LS), and linear transformation (LT). In all, 128 extrinsic normalization schemes were tested: 4 speaker adaptations x 4 frequency combinations

x 8 scales.

The CS method finds the mean frequency value ($\overline{F}$) across all transformed frequencies of all the vowels of a given speaker and subtracts this value from F_i':

$$F_i'' = F_i' - \overline{F}. \tag{2.7}$$

Nearey (1978) used the CS method with F_1 and F_2 in the constant log interval hypothesis (CLIH) method. Assmann, Nearey and Hogan (1982) obtained 84% accuracy using the CLIH method and LDA with the U method on 10 vowels spoken by 5 male and 5 female speakers.

While the CS method has the advantage of simplicity, the results of Fant (1966, 1975) suggest that F_1 and F_2 have different scalings. The CSi method extends the CS method by computing the centroid ($\overline{F_i}$) for each transformed frequency and subtracting this value from F_i':

$$F_i'' = F_i' - \overline{F_i}. \tag{2.8}$$

The CLIH2 method (Nearey, 1978), and CLIH3 method (Assmann et al., 1982), corresponding to the use of two (F_1, F_2) or three formants (F_1, F_2, F_3), are functionally equivalent to the CSi method in a log vowel space. Assmann, Nearey, and Hogan (1982) obtained 91% for CLIH2 and 93% for CLIH3 using LDA with the U method on the speakers described above.

The linear scale (LS) approach (Gerstman, 1968) finds the minimum and maximum frequency values for each F_i' across all vowels of a given speaker, then rescales each frequency to the range [0,999]:

$$F_i'' = 999 * (F_i' - F_i^{min'})/(F_i^{max'} - F_i^{min'}). \tag{2.9}$$

Gerstman hereby obtained 97.5% on the Peterson-Barney database using a metric derived from the database itself, so that training and testing occur on the same data set.

In the LT method (Hindle, 1978; Watrous, 1993; Zahorian & Jagharghi, 1991), a speaker-specific linear transformation matrix $\mathcal{A}$ transforms each speaker's frequencies into some prototypical frequency values. New frequencies are linear combinations of the original transformed frequencies:

$$F_i'' = \sum_{k=0}^{3} \alpha_{ik} F_k' + \beta_i. \tag{2.10}$$

The matrix $\mathcal{A}$ is derived using the least mean squares (LMS) algorithm (Widrow & Stearns, 1985) to minimize the mean squared error between a given speaker's fundamental and formant frequencies and the mean fundamental and formant frequencies across all speakers for each vowel. Hindle (1978) found that the LT method gave better performance than CS using the mean male, female, and child formant frequencies from the Peterson-Barney database. Zahorian and Jagharghi (1991), using the LT method, obtained 79.0% identification using a Bayesian maximum likelihood classifier after training on 11 vowels from a given speaker. The database they used consisted of the first three formants of 11 vowels, in 9 CVC contexts, spoken by 10 male, 10 female, and 10 child speakers. Watrous (1993), using a second-order back-propagation neural network with the LT normalization method, achieved 93.2% using only F_1 and F_2 on the Peterson-Barney database. However, Watrous (1993) did not use a separate set of speakers for training and testing.

Other normalization methods (Wakita, 1977; Bladon, Henton, & Pickering, 1984) require greater knowledge than specified in the Peterson-Barney database, such as knowledge of the spectral or temporal characteristics of a vowel.

Number	Arpabet	IPA	/hVd/
1	IY	/i/	heed
2	IH	/I/	hid
3	EH	/ɛ/	head
4	AE	/æ/	had
5	AH	/ʌ/	hud
6	AA	/ɑ/	hod
7	AO	/ɔ/	hawed
8	UH	/ʊ/	hood
9	UW	/u/	who'd
10	ER	/ɚ/	heard

Table 2.2: Vowels used in the Peterson and Barney (1952) study.

2.4 Peterson-Barney vowel database

Watrous (1991) recompiled Peterson and Barney's original data, which had proliferated into several inconsistent versions. Peterson and Barney tape recorded vowel data from 76 speakers (33 males, 28 females, 15 children), each speaking 10 vowels twice in a /hVd/ context (Table 2.2). Each vowel was analyzed during the steady-state portion to obtain the frequency values of the first three formants, F_1, F_2, F_3, as well as the fundamental frequency, F_0. This yielded a total of 1520 vowel tokens (76 speakers x 10 vowels x 2 repetitions).

For the present study, the Peterson-Barney database was split into a training and test set. Vowels spoken by approximately 30% of the speakers (10 male, 9 female, and 5 children) were randomly chosen to comprise the training set, with 1040 vowels spoken by the remaining speakers comprising the test set. The recognition task was thus far more challenging than one in which test speakers were also part of the training set.

2.5 Algorithms for comparing normalization methods

2.5.1 Fuzzy ARTMAP

Fuzzy ARTMAP (Figure 2·2) is a self-organizing neural network algorithm for adaptive categorization and prediction of nonstationary databases (Carpenter, Grossberg, & Reynolds, 1991a; Carpenter et al., 1992). During supervised training, the system learns to map (transformed) frequency vectors to 10 vowel categories. ARTMAP clusters frequency vectors on-line in one module (ART_a) and vowel categories in a second module (ART_b). An intervening map field (F^{ab}) adaptively associates frequency categories to vowel categories. The main components of the fuzzy ARTMAP system will now be outlined.

The ART_a and ART_b modules cluster the input vector and output vector, respectively. An input $\mathbf{a}$ to ART_a field F_0^a is an M-dimensional vector with component values between 0 and 1. Complement coding (Carpenter, Grossberg, & Rosen, 1991b) enables the system to encode both absent and present features. After complement coding, the 2M-dimensional vector $\mathbf{A} \equiv (\mathbf{a}, \mathbf{a}^c)$ becomes the input to F_1^a, where $a_i^c \equiv 1 - a_i$. Activity at F_1^a activates a category node J at F_2^a. This active F_2^a category sends top-down signals to F_1^a, where internal system dynamics determine whether the match between the bottom-up input and the top-down learned weight vector $\mathbf{w}_J^a$ is good enough to permit learned weight changes. The matching, or vigilance, criterion is satisfied if

$$\rho_a |\mathbf{A}| < |\mathbf{A} \wedge \mathbf{w}_J|, \tag{2.11}$$

where $\rho_a \in [0, 1]$ is a dimensionless parameter called vigilance, and where $\wedge$ represents the fuzzy AND, or component-wise minimum, operation. If the match does not meet the vigilance criterion, category J is reset and another category node in F_2^a

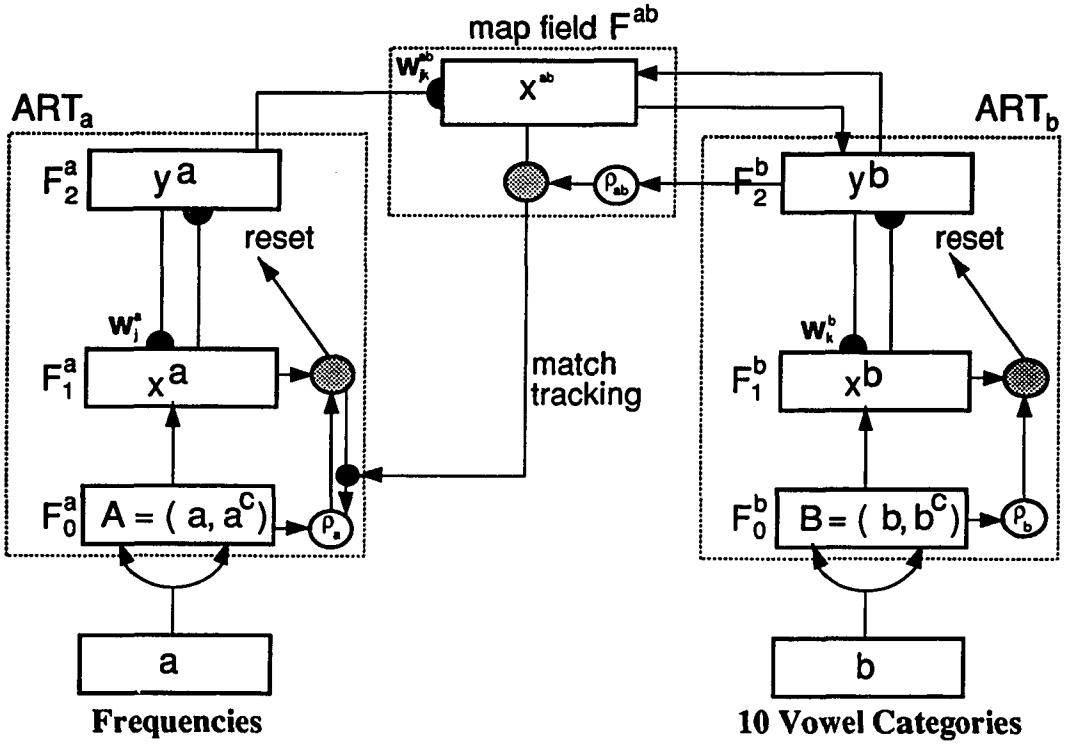


Figure 2·2: Fuzzy ARTMAP architecture with its three component modules: ART_a, ART_b, and the map field F^{ab}. The ART_a module transforms the M-dimensional input vector $\mathbf{a}$ into a 2M-dimensional input vector $\mathbf{A} = (\mathbf{a}, \mathbf{a}^c)$ at the F_0^a field through complement coding. $\mathbf{A}$ is the input to the ART_a field F_1^a. Similarly, the input to the ART_b field F_1^b is the vector $\mathbf{B} = (\mathbf{b}, \mathbf{b}^c)$. The activation of field F_1^a causes a category to become active at F_2^a, which in turn leads to a prediction by ART_a at the map field F^{ab}. If the prediction is disconfirmed at ART_b, inhibition of the map field activity starts the match tracking process: the ART_a vigilance ρ_a is raised slightly over the match ratio. This in turn causes an ART_a search which leads to either the activation of another existing ART_a category that correctly predicts $\mathbf{B}$, or a new uncommitted ART_a category node.

is selected. This search process continues until an active F_2^a node meets the match criterion (2.11), or a new category is selected.

During fuzzy ARTMAP training, the input pattern (**a**) and output pattern (**b**) select an F_2^a category node J and an F_2^b category node K, respectively. The map field F^{ab} associates the two categories, unless J had previously learned to predict a different F_2^b category $\hat{K}$. When such a predictive mismatch occurs, another F_2^a node is chosen through a fuzzy ARTMAP control process called *match tracking.* During testing, an input pattern **a** presented to ART_a activates a category in ART_b via the map field F_{ab}. The chosen output pattern **b** then constitutes the test set prediction.

2.5.2 K-Nearest Neighbor

The K-Nearest Neighbor (K-NN) algorithm (Duda & Hart, 1973; Dasarathy, 1991) finds, for each test point, the K nearest training points, with distance measured by some metric. The vowel categories for the K neighbors are tallied, and the test point is assigned to the vowel category with the largest number of votes. If a tie occurs between two or more vowel categories, the category with the minimum total distance is chosen. In the simulations, two different metrics, city block (L_1) and Euclidean (L_2), were compared. The L_1 norm, also used in fuzzy ARTMAP, equals the sum of the absolute values of the differences between the vector components:

$$\|\mathbf{x} - \mathbf{y}\|_1 = \sum_i |x_i - y_i|. \tag{2.12}$$

The L_2 norm, or Euclidean metric, is defined by:

$$\|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_i (x_i - y_i)^2}. \tag{2.13}$$

2.6 Comparative evaluation of speaker normalization methods

Preliminary simulations on different normalization methods were run to select parameters for the K-NN and fuzzy ARTMAP recognition systems. For fuzzy ARTMAP, performance on the test set was evaluated after the system had achieved 100% correct performance on the training set. For the K-NN systems, the number of neighbors (K) was fixed at 10 throughout. Performance trends were fairly insensitive to system parameters.

2.6.1 Intrinsic methods

Fuzzy ARTMAP and K-NN evaluations of the 32 intrinsic methods are shown in Tables 2.3 and 2.4, respectively, and in Figure 2·3. Among the psychophysical measures, the bark and bark with end-correction perform slightly better than mel or ERB scale, while there was little difference among the three log measures. Both the psychophysical and log measures showed a slight advantage over the nonscaled formants. The Diff methods increased the performance of the psychophysical measures while decreasing the log measures. In fact, for all three classifiers, the best performance for the log scales was achieved with $[F_0', F_1', F_2', F_3']$. However, the overall best intrinsic normalization method is the Diff All method using bark or bark with end-correction scaled formants. Fuzzy ARTMAP and L_1 K-NN achieved 83.1% and 85.5%, respectively, using bark scale; and L_2 K-NN achieved 85.8% using bark scale with end correction, just edging out bark scale (85.5%).

2.6.2 Extrinsic methods

In general, the LT method performed better than the other extrinsic schemes, as follows.

Vowel Space		$[F_1', F_2']$		$[F_0', F_1', F_2', F_3']$		Diff Subset		Diff All
Scale	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)
N	1	66.4 (123.1)	9	78.4 (63.5)	17	80.4 (55.8)	25	80.7 (57.5)
B	2	66.0 (123.7)	10	79.1 (61.6)	18	81.4 (56.3)	26	**83.1** (43.9)
Be	3	65.8 (123.1)	11	78.6 (63.9)	19	80.8 (54.8)	27	**83.1** (43.4)
M	4	65.5 (124.3)	12	79.0 (62.2)	20	79.8 (57.1)	28	81.6 (46.3)
ERB	5	64.9 (124.8)	13	79.1 (62.3)	21	77.7 (66.1)	29	79.4 (49.4)
$log_{1.06}$	6	65.4 (122.0)	14	79.4 (60.7)	22	72.1 (73.2)	30	74.2 (58.9)
log_e	7	65.5 (121.9)	15	79.4 (60.6)	23	72.3 (72.5)	31	74.0 (58.8)
log_{10}	8	65.5 (122.1)	16	79.4 (60.8)	24	71.9 (73.9)	32	74.2 (58.9)

Table 2.3: Fuzzy ARTMAP test set performance with intrinsic normalization. Numbers in parentheses give the average number of F_2^a nodes after training. Vowel space scales: N = nonscaled, B = bark scaled, Be = bark scaled with end-correction, M = mel scaled, ERB = equivalent rectangular bandwidth scaled, log_e = natural logarithm scaled, $log_{1.06}$ = semitone scaled, and log_{10} = log base 10 scaled. Intrinsic normalization methods 1-8 use only the first two formants $[F_1', F_2']$; methods 9-16 use $[F_0', F_1', F_2', F_3']$; for methods 17-32, differences between the transformed $F_0', \ldots, F_3'$ were computed. Methods 17-24 (Diff Subset) employ the three differences $F_1' - F_0', F_2' - F_1', F_3' - F_2'$; methods 25-32 (Diff All) employ all six differences $F_1' - F_0', F_2' - F_0', F_3' - F_0', F_2' - F_1', F_3' - F_1', F_3' - F_2'$. Fuzzy ARTMAP simulation parameters: $\bar{\rho}_a = 0.0$, $\alpha = 0.1$, $\beta = 1.0$.

Vowel Space	$[F_1', F_2']$		$[F_0', F_1', F_2', F_3']$		Diff Subset		Diff All	
Scale	Id	%	Id	%	Id	%	Id	%
			L_1 K-NN					
N	1	75.2	9	76.8	17	78.9	25	76.8
B	2	74.3	10	82.6	18	83.7	26	**85.5**
Be	3	74.3	11	81.4	19	84.1	27	**85.4**
M	4	74.6	12	82.0	20	83.4	28	82.9
ERB	5	73.8	13	83.5	21	82.1	29	82.1
$log_{1.06}$	6	74.5	14	82.0	22	76.1	30	77.2
log_e	7	74.5	15	82.0	23	76.0	31	77.3
log_{10}	8	74.5	16	82.1	24	76.0	32	77.2
			L_2 K-NN					
N	1	75.2	9	75.1	17	77.1	25	76.3
B	2	75.1	10	82.6	18	84.5	26	**85.5**
Be	3	75.1	11	83.1	19	84.0	27	**85.8**
M	4	75.3	12	82.4	20	83.0	28	82.5
ERB	5	74.9	13	82.7	21	81.4	29	81.9
$log_{1.06}$	6	74.8	14	82.5	22	76.1	30	77.1
log_e	7	74.8	15	82.5	23	76.3	31	77.1
log_{10}	8	74.8	16	82.5	24	76.0	32	77.1

Table 2.4: L_1 and L_2 K-NN (K = 10) test set performance with intrinsic normalization. The vowel space scales are specified in Table 2.3.

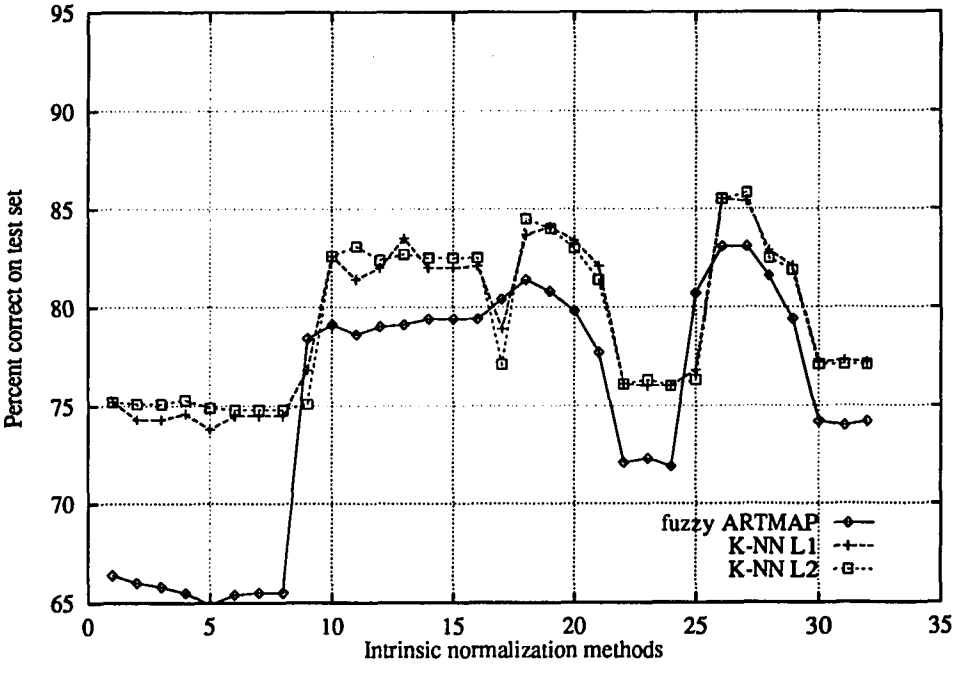


Figure 2·3: Comparison between fuzzy ARTMAP and K-NN for intrinsic normalization methods. Normalization identification numbers are in Tables 2.3 and 2.4. Methods 26 and 27 (B and Be, Diff All) have the best performance.

Vowel Space		$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$	Diff Subset		Diff All	
Scale	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)
N	1	66.2 (134.3)	9	79.8 (56.2)	17	77.4 (65.8)	25	79.1 (62.9)
B	2	78.7 (80.5)	10	81.2 (43.3)	18	81.4 (56.3)	26	**83.1** (43.9)
Be	3	78.6 (80.7)	11	82.1 (44.7)	19	80.8 (54.8)	27	**83.1** (43.4)
M	4	77.1 (82.3)	12	78.5 (46.3)	20	77.5 (61.9)	28	80.5 (51.9)
ERB	5	79.6 (73.9)	13	80.9 (45.7)	21	77.5 (64.4)	29	79.6 (49.3)
$log_{1.06}$	6	80.9 (66.6)	14	81.3 (49.1)	22	75.7 (64.9)	30	77.6 (50.1)
log_e	7	80.9 (65.8)	15	81.3 (49.1)	23	75.7 (64.9)	31	77.6 (50.1)
log_{10}	8	80.9 (66.5)	16	81.3 (49.1)	24	75.7 (64.7)	32	77.6 (50.0)

Table 2.5: Fuzzy ARTMAP test set performance with centroid subtraction across all frequencies (CS) extrinsic normalization.

The results for the centroid subtraction across all frequency (CS) extrinsic method using fuzzy ARTMAP and K-NN are shown in Tables 2.5 and 2.6, respectively, and summarized graphically in Figure 2·4. For the CS method, the log measures perform better for the $[F_1'', F_2'']$ case; but otherwise, the psychophysical measures again perform better. The best performance for fuzzy ARTMAP was achieved using bark Diff All (83.1%); and ERB $[F_0'', F_1'', F_2'', F_3'']$ for the L_1 (87.3%) and L_2 (86.9%) K-NN.

The results for the CSi extrinsic method using fuzzy ARTMAP and K-NN are shown in Tables 2.7 and 2.8, respectively, and in Figure 2·5. Here ARTMAP and K-NN disagree as to which intrinsic method is best. Fuzzy ARTMAP performs optimally for bark with end correction Diff All (88.1%); and the K-NN methods perform optimally for the log $[F_0'', F_1'', F_2'', F_3'']$ intrinsic method (90.6% for L_1 and 90.9% for L_2).

The results for the LS extrinsic method using fuzzy ARTMAP and K-NN are shown in Tables 2.9 and 2.10, respectively, and summarized in Figure 2·6. For LS normalization, the bark Diff All method was nearly the best for both fuzzy ARTMAP (84.8%) and the K-NN (88.8%), although the best performance for L_1 K-NN was

Vowel Space	$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$		Diff Subset		Diff All	
Scale	Id	% Correct	Id	% Correct	Id	% Correct	Id	% Correct
				L_1 K-NN				
N	1	71.7	9	77.9	17	79.6	25	76.9
B	2	83.8	10	86.1	18	83.7	26	85.5
Be	3	83.8	11	86.3	19	84.1	27	85.4
M	4	84.7	12	86.0	20	82.1	28	82.7
ERB	5	85.4	13	**87.3**	21	81.2	29	82.7
$log_{1.06}$	6	86.1	14	85.3	22	81.0	30	81.8
log_e	7	86.1	15	85.3	23	81.0	31	81.8
log_{10}	8	86.1	16	85.4	24	80.9	32	81.8
				L_2 K-NN				
N	1	73.3	9	77.3	17	78.6	25	75.5
B	2	84.2	10	86.3	18	84.5	26	85.5
Be	3	84.2	11	86.3	19	84.0	27	85.8
M	4	84.5	12	86.0	20	82.8	28	83.0
ERB	5	85.3	13	**86.9**	21	82.0	29	83.3
$log_{1.06}$	6	86.3	14	85.8	22	80.3	30	81.3
log_e	7	86.3	15	85.8	23	80.3	31	81.3
log_{10}	8	86.3	16	85.8	24	80.3	32	81.3

Table 2.6: L_1 and L_2 K-NN test set performance with centroid subtraction across all frequencies (CS) extrinsic normalization.

Vowel Space	$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$		Diff Subset		Diff All	
Scale	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)
N	1	81.1 (68.1)	9	83.5 (37.1)	17	83.4 (40.9)	25	84.5 (35.3)
B	2	84.5 (53.9)	10	86.1 (33.9)	18	87.0 (37.5)	26	87.6 (28.2)
Be	3	84.5 (53.5)	11	86.2 (35.1)	19	86.4 (40.5)	27	**88.1** (29.1)
M	4	84.8 (58.6)	12	86.1 (33.1)	20	86.9 (34.7)	28	87.7 (28.3)
ERB	5	85.3 (55.5)	13	86.4 (31.9)	21	86.3 (35.0)	29	87.4 (27.5)
$log_{1.06}$	6	86.0 (55.2)	14	86.5 (32.8)	22	85.2 (37.8)	30	86.7 (29.0)
log_e	7	86.0 (55.5)	15	86.5 (32.8)	23	85.1 (37.8)	31	86.8 (28.9)
log_{10}	8	85.9 (55.2)	16	86.5 (32.9)	24	85.0 (37.8)	32	86.7 (29.0)

Table 2.7: Fuzzy ARTMAP test set performance with centroid subtraction for each frequency (CSi) extrinsic normalization.

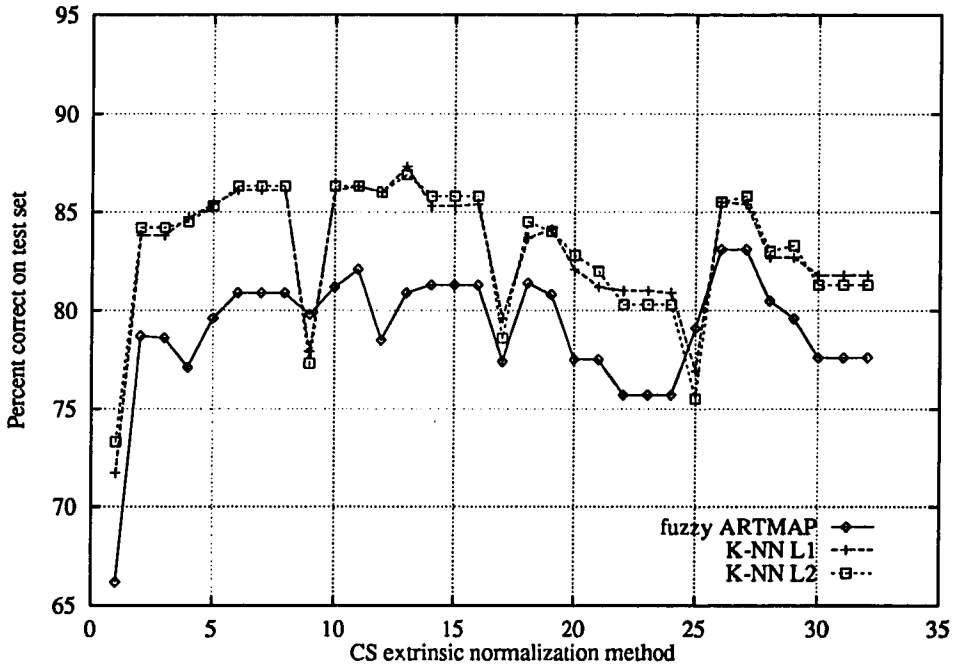


Figure 2·4: Comparison between fuzzy ARTMAP and K-NN for the centroid subtraction across all frequencies (CS) extrinsic normalization method (Tables 2.5 and 2.6). Fuzzy ARTMAP performed best with methods 26 and 27 (B and Be, Diff All), while K-NN performed best with method 13 (ERB, $[F_0'', F_1'', F_2'', F_3'']$).

Vowel Space	$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$		Diff Subset		Diff All	
Scale	Id	% Correct	Id	% Correct	Id	% Correct	Id	% Correct
				L_1 K-NN				
N	1	83.1	9	83.8	17	84.5	25	83.3
B	2	87.4	10	89.3	18	88.4	26	88.6
Be	3	87.3	11	88.8	19	88.0	27	88.6
M	4	87.2	12	89.7	20	89.7	28	89.3
ERB	5	88.5	13	90.0	21	90.5	29	90.0
$log_{1.06}$	6	88.6	14	**90.6**	22	90.0	30	89.2
log_e	7	88.6	15	**90.6**	23	90.0	31	89.2
log_{10}	8	88.5	16	**90.6**	24	90.0	32	89.2
				L_2 K-NN				
N	1	83.4	9	83.0	17	84.3	25	83.0
B	2	86.9	10	88.8	18	88.1	26	88.8
Be	3	86.9	11	89.1	19	87.9	27	88.3
M	4	87.3	12	89.3	20	89.4	28	89.5
ERB	5	88.8	13	90.0	21	89.6	29	89.9
$log_{1.06}$	6	88.5	14	**90.9**	22	89.7	30	89.4
log_e	7	88.5	15	**90.9**	23	89.7	31	89.4
log_{10}	8	88.5	16	**90.9**	24	89.7	32	89.4

Table 2.8: L_1 and L_2 K-NN test set performance with centroid subtraction for each frequency (CSi) extrinsic normalization.

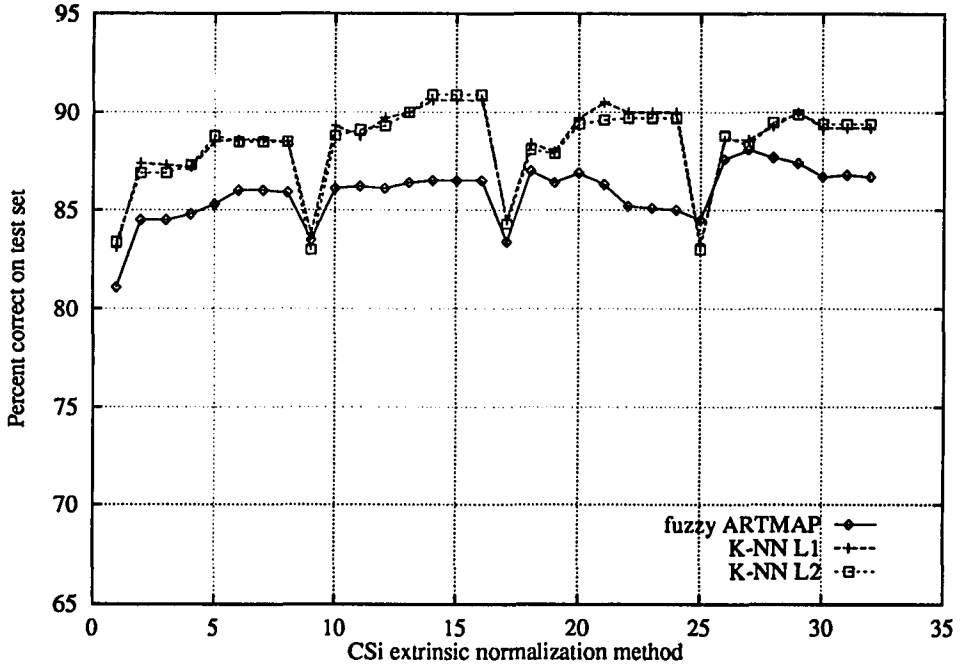


Figure 2.5: Comparison between fuzzy ARTMAP and K-NN for the centroid subtraction for each frequency (CSi) extrinsic normalization method (Tables 2.7 and 2.8). Fuzzy ARTMAP performed best with method 26 (B, Diff All), and K-NN performed best with methods 14 through 16 ($log_e/log_{1.06}/log_{10}$, $[F_0'', F_1'', F_2'', F_3'']$).

Vowel Space	$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$		Diff Subset		Diff All	
Scale	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)
N	1	79.1 (82.6)	9	81.5 (52.3)	17	84.0 (49.6)	25	83.0 (35.4)
B	2	77.7 (86.1)	10	81.7 (50.1)	18	84.4 (46.9)	26	**84.8** (30.1)
Be	3	77.6 (86.1)	11	81.4 (54.3)	19	84.0 (50.6)	27	**84.8** (31.5)
M	4	78.2 (86.7)	12	81.9 (50.7)	20	83.7 (49.0)	28	84.4 (31.7)
ERB	5	77.5 (87.8)	13	81.7 (51.9)	21	82.9 (51.0)	29	84.9 (32.1)
$log_{1.06}$	6	76.5 (92.3)	14	81.7 (51.6)	22	82.0 (54.3)	30	82.7 (35.7)
log_e	7	76.5 (92.3)	15	81.7 (51.6)	23	82.0 (54.3)	31	82.7 (35.7)
log_{10}	8	76.5 (92.3)	16	81.7 (51.6)	24	82.0 (54.3)	32	82.7 (35.7)

Table 2.9: Fuzzy ARTMAP test set performance with linear scale (LS) extrinsic normalization.

ERB Diff All (89.0%), and bark Diff Subset (89.0%) for L_2 K-NN. Also, the psychophysical measures perform slightly better than the log measures.

The results for the LT extrinsic method using fuzzy ARTMAP and K-NN are shown in Tables 2.11 and 2.12, respectively, and in Figure 2·7. Comparing Figure 2·7 with Figures 2·4-2·6, the other extrinsic methods, shows that LT extrinsic normalization has the best performance. Once again, the psychophysical measures perform slightly better than the log measures. However, fuzzy ARTMAP and the K-NN disagree on which psychophysical measure is best. For fuzzy ARTMAP, the bark Diff All method achieves 92.2%. For the K-NN algorithms, the $[F_0'', F_1'', F_2'', F_3'']$ methods perform best, with L_1 K-NN achieving 94.3% with the mel scale, and L_2 K-NN achieving 94.6% with the ERB scale.

2.7 Discussion

2.7.1 Fuzzy ARTMAP vs. K-NN

While having similar tendencies, the K-NN algorithms tended to outperform fuzzy ARTMAP. However, the improved performance achieved by K-NN comes at a cost of storing all 480 training points. Fuzzy ARTMAP coded between 22 and 135 F_2^a nodes,

Vowel Space		$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$	Diff Subset		Diff All	
Scale	Id	% Correct	Id	% Correct	Id	% Correct	Id	% Correct
				L_1 K-NN				
N	1	82.8	9	83.8	17	87.7	25	86.5
B	2	81.9	10	85.8	18	88.1	26	88.8
Be	3	82.0	11	84.3	19	88.1	27	88.3
M	4	82.4	12	84.8	20	87.9	28	88.5
ERB	5	82.1	13	85.2	21	88.4	29	**89.0**
$log_{1.06}$	6	81.8	14	84.4	22	87.5	30	87.7
log_e	7	81.8	15	84.4	23	87.5	31	87.7
log_{10}	8	81.8	16	84.4	24	87.5	32	87.7
				L_2 K-NN				
N	1	82.7	9	81.3	17	88.4	25	86.3
B	2	82.1	10	83.8	18	**89.0**	26	88.8
Be	3	82.0	11	82.4	19	87.8	27	88.0
M	4	81.5	12	83.8	20	88.8	28	88.2
ERB	5	82.3	13	83.1	21	88.2	29	87.9
$log_{1.06}$	6	81.5	14	83.1	22	87.8	30	87.0
log_e	7	81.5	15	83.1	23	87.8	31	87.0
log_{10}	8	81.5	16	83.1	24	87.8	32	87.0

Table 2.10: L_1 and L_2 K-NN test set performance with linear scale (LS) extrinsic normalization.

Vowel Space		$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$	Diff Subset		Diff All	
Scale	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)	Id	% (F_2^a)
N	1	89.5 (41.7)	9	89.4 (29.1)	17	90.7 (29.5)	25	90.9 (23.9)
B	2	88.7 (44.2)	10	91.0 (26.1)	18	91.2 (27.9)	26	**92.2** (22.0)
Be	3	88.7 (43.8)	11	90.4 (27.3)	19	90.8 (28.8)	27	91.6 (22.5)
M	4	88.9 (42.7)	12	89.4 (26.1)	20	91.2 (28.7)	28	91.8 (26.1)
ERB	5	88.6 (42.3)	13	90.3 (25.5)	21	90.5 (29.0)	29	91.6 (22.5)
$log_{1.06}$	6	86.2 (54.9)	14	88.5 (27.8)	22	88.4 (33.1)	30	89.7 (28.2)
log_e	7	88.2 (49.1)	15	87.8 (33.9)	23	88.9 (31.3)	31	90.7 (25.2)
log_{10}	8	87.2 (57.8)	16	82.6 (42.5)	24	88.5 (33.5)	32	90.3 (25.1)

Table 2.11: Fuzzy ARTMAP test set performance with linear transformation (LT) extrinsic normalization.

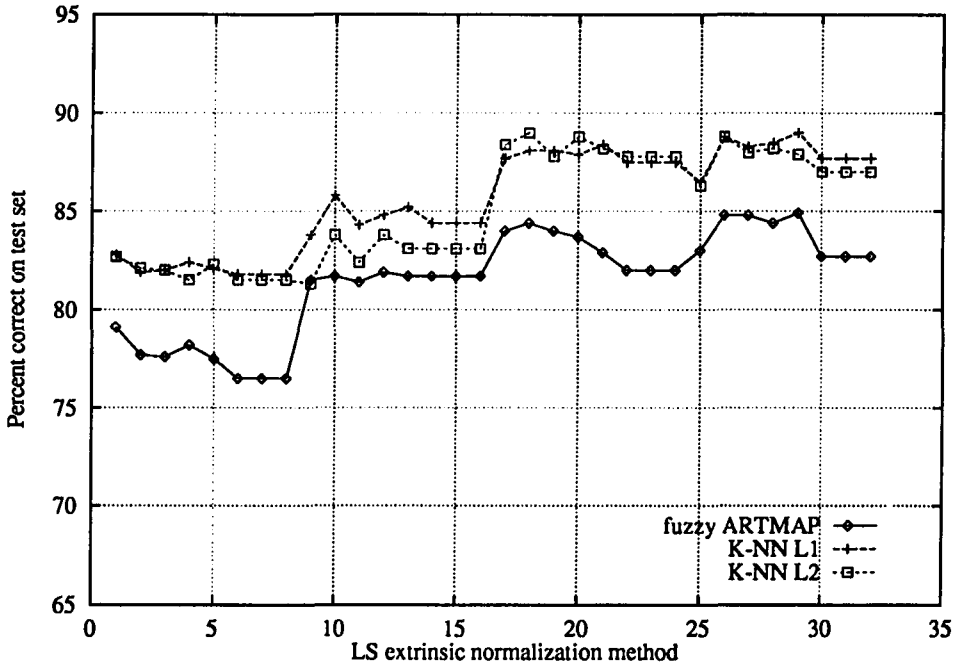


Figure 2·6: Comparison between fuzzy ARTMAP and K-NN for the linear scale (LS) extrinsic normalization method (Table 2.9 and 2.10). Fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN performed best with methods 26 and 27 (B and Be, Diff All), method 29 (ERB, Diff All), and method 18 (B, Diff Subset), respectively.

Vowel Space	$[F_1'', F_2'']$		$[F_0'', F_1'', F_2'', F_3'']$		Diff Subset		Diff All	
Scale	Id	% Correct	Id	% Correct	Id	% Correct	Id	% Correct
				L_1 K-NN				
N	1	92.0	9	91.4	17	92.0	25	92.2
B	2	91.5	10	94.0	18	93.0	26	93.5
Be	3	91.6	11	93.7	19	93.1	27	93.3
M	4	92.6	12	**94.3**	20	93.4	28	92.7
ERB	5	91.0	13	94.2	21	93.3	29	93.0
$log_{1.06}$	6	88.4	14	93.3	22	92.1	30	92.4
log_e	7	90.7	15	92.1	23	92.6	31	93.1
log_{10}	8	89.5	16	87.1	24	92.4	32	92.7
				L_2 K-NN				
N	1	91.7	9	90.1	17	92.2	25	92.1
b	2	91.7	10	94.1	18	93.1	26	93.7
Be	3	91.8	11	94.0	19	92.8	27	93.3
M	4	92.0	12	94.2	20	93.6	28	92.6
ERB	5	91.0	13	**94.6**	21	93.1	29	93.4
¿$log_{1.06}$	6	88.5	14	93.4	22	91.7	30	92.4
log_e	7	91.0	15	91.6	23	92.8	31	92.9
log_{10}	8	89.5	16	88.3	24	92.0	32	92.9

Table 2.12: L_1 and L_2 K-NN test set performance with linear transformation (LT) extrinsic normalization.

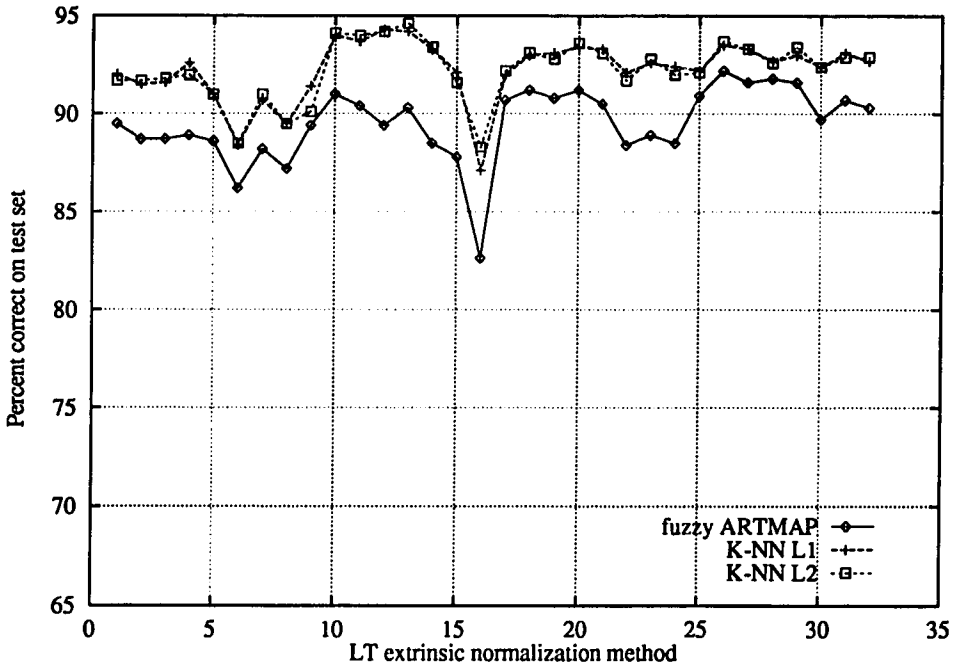


Figure 2·7: Comparison between fuzzy ARTMAP and K-NN for the linear transformation (LT) extrinsic normalization method (Tables 2.11 and 2.12). Fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN performed best with methods 26 (B, Diff All), method 12 (M, $[F_0'', F_1'', F_2'', F_3'']$), and method 13 (ERB, $[F_0'', F_1'', F_2'', F_3'']$), respectively.

which provides a compression of 3.5 to 21.8 compared to the storage requirements of K-NN. Similar savings are achieved in computation time during performance. These differences in storage and computation can be a major factor in large-scale applications.

2.7.2 Differences between vowel space scales

The results from both the intrinsic and most extrinsic methods show that both psychophysical (B, Be, M, ERB) and log transformations are better than none (N). There was little difference between Bark (B) and bark with end correction (Be). Similarly, the three log measures show no major differences among themselves, except in the LT extrinsic method, where the natural log measure performs better for all except $[F_0'', F_1'', F_2'', F_3'']$, where the semitone scale performs better. The performance of fuzzy ARTMAP, for the intrinsic and all the extrinsic methods, was optimal for either bark or bark with end correction, with Diff All. While the K-NN algorithms varied more, these methods chose the psychophysical measures for all but the CSi method. Thus, on the whole, the psychophysical measures provide better speaker-independent representation than log measures.

2.7.3 Intrinsic methods

For intrinsic methods, bark differences are usually the best speaker normalization method with the best performance achieved by bark Diff All, with 83.1% for ARTMAP and 85.5% for L_1 K-NN; and bark with end correction Diff All with 85.8% for L_2 K-NN.

The performance for log measures using ratios was about 5% less than the performance using $[F_0', F_1', F_2', F_3']$. Thus, speaker normalization methods using logs of formant ratios seems to be a poorer invariant representation than the simpler method

of converting the frequencies to a log scale.

2.7.4 Extrinsic methods

Among the extrinsic normalization schemes, the LT method performs best, followed by CSi, LS, and CS. The LT method works best using either the bark Diff All method or ERB/mel transformed $[F_0'', F_1'', F_2'', F_3'']$. The second best extrinsic method is the CSi method using either bark with end-correction Diff All or log $[F_0'', F_1'', F_2'', F_3'']$. The LS method with bark Diff All/ERB Diff Subset proved the next best, followed by the CS method with either bark Diff All or ERB $[F_0'', F_1'', F_2'', F_3'']$.

While LT performs best it requires the most *a priori* knowledge, namely *labeled* training set data points. As a model of human vowel perception, the LT method seems unlikely since the listener would have to identify a speaker's vowels ahead of time in order to create the transformation matrix, which is needed to identify the vowels. However, for a machine recognition application, wherein a speaker can state a specified utterance allowing the machine to create the transformation matrix, the LT method seems feasible. On the other hand, the CSi method, which has the same complexity as CS or LS, performs almost as well and does not require the identity of vowels for the speaker adaptation.

2.8 Summary

This research has developed a method for comparing a large number of normalization methods in a speaker independent fashion that is fast and systematic, and that is readily applicable to other areas. In addition, one sees that both fuzzy ARTMAP and K-NN had similar trends with K-NN performing better but requiring ten times as much memory. In comparing psychophysical and log measures, one generally sees that the psychophysical measures outperform the log measures: K-NN performed

best using psychophysical scales for most normalization methods; fuzzy ARTMAP performed optimally using bark scale, a psychophysical measure, with all possible differences (Diff All), for all the normalization methods. Thus, bark scale using Diff All seems best for creating a more canonical representation for easier classification.

Chapter 3

A psychophysical study of adaptation to silent intervals during variable-rate speech

3.1 Introduction

Researchers in speech perception have been searching for cues in the speech signal that invariantly identify linguistic units; e.g. phonemes. These invariant cues specify the phoneme across contexts, speaking rates, and speakers. However, this search has not been fruitful, and instead it has led to the finding that there are many perceptually relevant cues that influence the identification of phonemes, and that these cues can vary with context, rate, and can trade against each other.

Some of the acoustic cues that listeners employ to distinguish phonemes include formant frequency transitions at the onset and offset of voiced portions (Liberman, Delattre, & Gerstman, 1954; Sharf & Ohde, 1984), duration of voiced portions [1] (Ainsworth, 1972; Just, Suslick, Michaels, & Shockey, 1978; Miller & Baer, 1983), duration of noise spectra (Repp, Lieberman, Eccardt, & Pesetsky, 1978), and duration of closure intervals (Dorman & Raphael, 1980). Formant frequencies are frequencies in the speech signal where there is greater energy; and closure interval is the time period during which a closure in the vocal tract is produced, and as a consequence there is a silent interval (but usually, there is also very low frequency energy present). For example, Figure 3·1 shows a spectrogram of the intervocalic

[1] The voiced portion corresponds to the voiced formant transitions and the adjacent vowel.

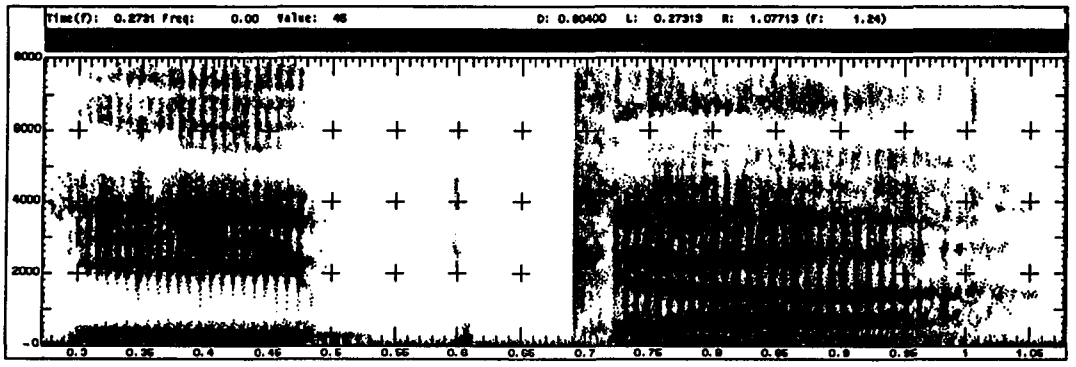


Figure 3·1: Spectrogram of intervocalic stop consonant /ada/ uttered by a male speaker.

stop /ada/. In the figure, the formant frequencies are shown as the darker horizontal lines, with the closure interval occurring between 0.48 and 0.69 seconds. The voiced portion before the closure interval is referred to as the vowel-consonant (VC) transition, and the portion after the closure is the consonant-vowel (CV) transition.

3.1.1 Cue adaptation

These different types of durations are used as cues for phonetic identification and also for detecting word boundaries in speech (Repp et al., 1978; Repp, 1980; Cutler & Butterfield, 1990). Since these duration cues vary as a function of speaking rate and stress (Gay, 1978; Repp et al., 1978; Port, 1979; Tartter, Kat, Samuel, & Repp, 1983), the distribution of these intervals can influence how speech sounds are categorized for purposes of recognition. Thus, the listener has to adapt to the distribution and compensate for rate effects. Repp (1980) investigated a listener's ability to adapt to silent intervals in stop consonant clusters.Stop consonant clusters are speech segments consisting of two adjacent stop consonants, e.g. /ad-ga/. Repp found that by skewing the distribution of silent intervals in the experiment, the psychometric curve for hearing only one or both stop consonants shifted close to the mean of the

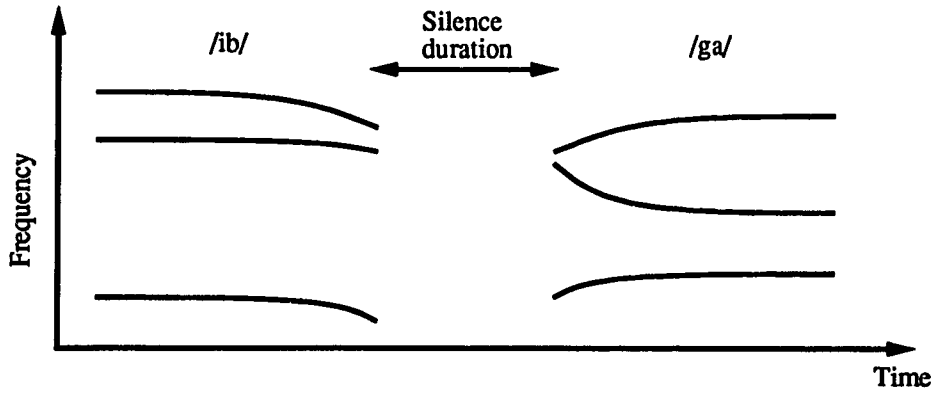


Figure 3·2: Schematic representation of the /ib-ga/ cluster token from Repp (1980).

skewed distribution, as well as altering its slope.

3.1.2 Repp (1980) experiment

Repp (1980) varied the silence duration between two stop consonants in a synthetic $/VC_1 - C_2V/$ syllable, and had subjects state whether they heard one or two stop consonants. The experiment consisted of two cases: the cluster case had formant transitions corresponding to 2 different stops, $C_1 \neq C_2$; and the geminate case had transitions corresponding to the same stop, $C_1 = C_2$. For both the cluster and geminate case, the VC_1 was /ib/. The C_2V was /ga/ for the cluster condition and /ba/ for the geminate condition. Figure 3·2 shows a schematic representation of a cluster token. In the cluster case, the subject circled "bg" if they heard both stop consonants, or they circled "g" if they heard only one stop consonant. In the geminate case the subject responded similarly by circling "bb" if they heard two stop consonants, or "b" if they heard only one stop consonant.

Each case had three conditions: no anchor, low anchor, and high anchor condition. These conditions correspond to how the silence durations were distributed over the condition. The distributions for the cluster case is shown in Figure 3·3. In the

no anchor condition there were 9 tokens of 11 silence durations, and thus, an equal distribution. In the low anchor condition, there were 30 tokens of the shortest silence duration (15ms), and 10 tokens each of the next 7 silence durations. Thus, the low anchor condition had a skewed distribution with more shorter silence durations than longer durations. The high anchor case is the opposite of the low anchor condition, with 30 tokens of the longest silence duration (115ms), and 10 tokens each of the previous 7 silence durations.

By presenting these different distributions, Repp found that the psychometric curve for hearing one versus two stops shifted and changed its slope as a function of the range of silence intervals and the number of tokens of each silent interval. Figure 3·4 shows the results that Repp obtained by averaging the results of the identification curves over 8 subjects. Note that the slope of the curve was more shallow for the no anchor case than for either the anchored cases. Since Repp pooled the response across subjects, the resulting shift and slope change in the curves could be due to averaging.

This paper replicates and extends the Repp (1980) cluster experiment to determine if the shift and slope changes in the response curves are due to pooling across subjects, and what the underlying mechanism is in silent interval adaptation. The first experiment replicates the cluster condition of Repp (1980), and finds that the individual subject data, while quite varied across subjects, show the curve shift as a function of the silence distribution. In addition, the pooled data shows a slope change for the no anchor condition. The second experiment extends the first experiment by performing the experiment with a different stop consonant cluster, /ad-ga/, showing the shifts in the curves are valid, with no apparent slope change, and having less variability across subjects. The third and fourth experiments investigate the adaptation mechanism. The third experiment investigates the influence of the overall variance

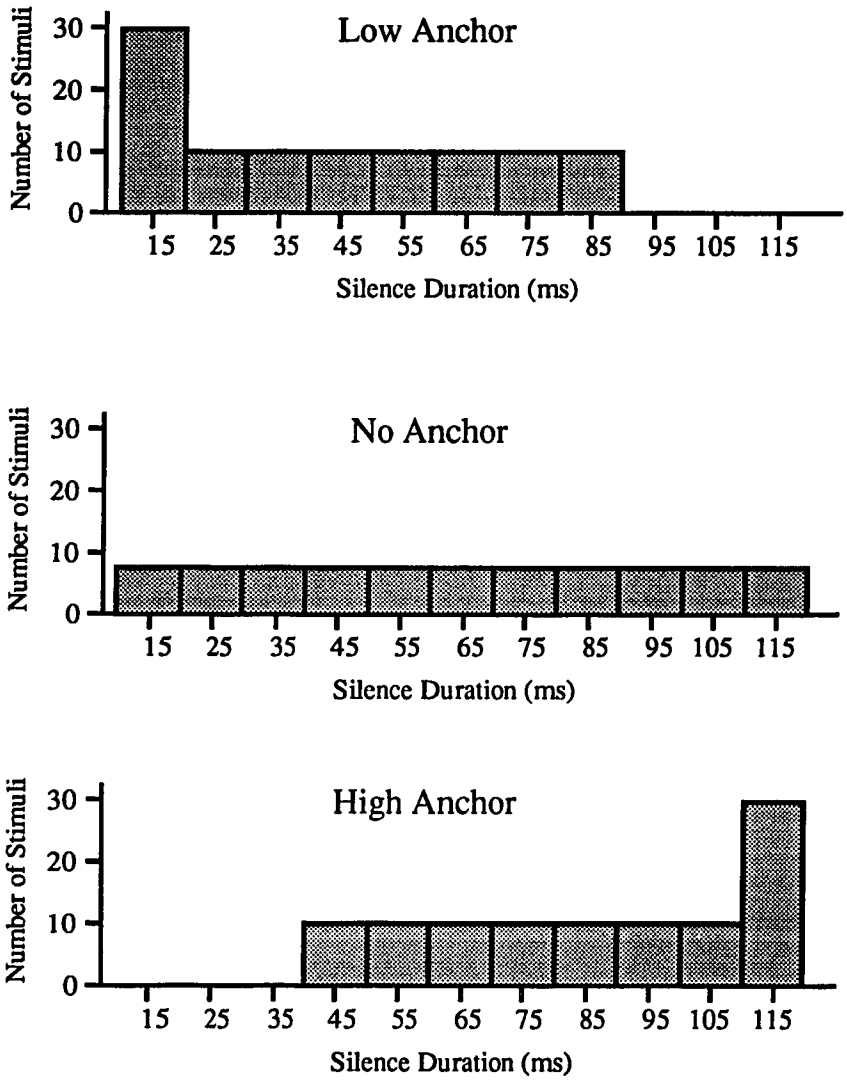


Figure 3-3: Distribution of the silent intervals for the three anchor conditions for the cluster case of Repp (1980).

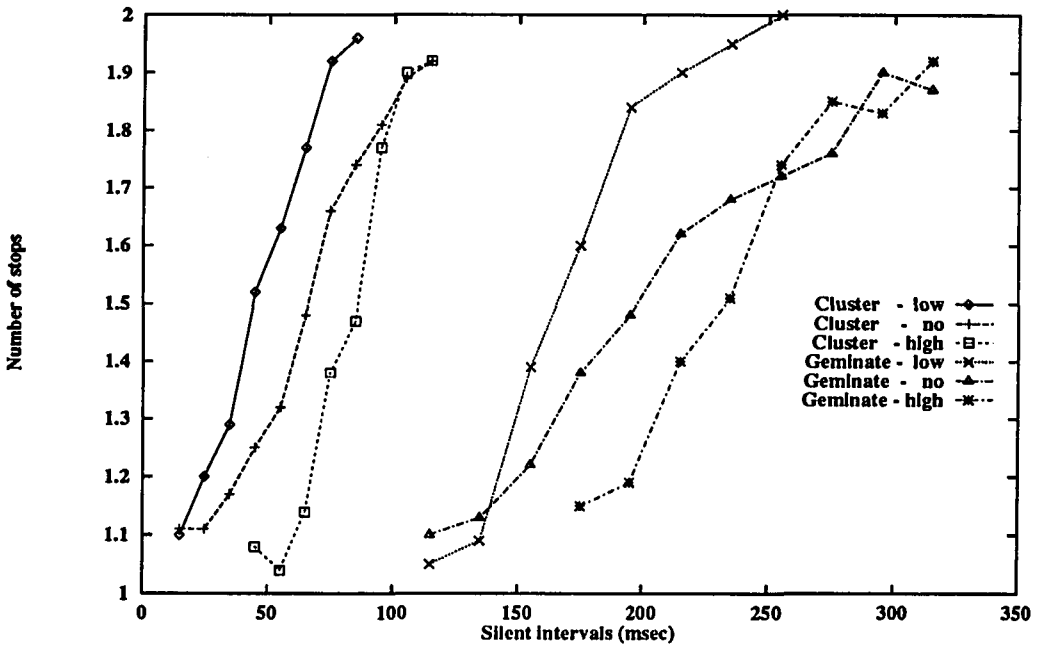


Figure 3·4: Results from the Repp (1980) experiment for the three anchor conditions for both the cluster and geminate case as averaged over 8 subjects. The datapoints were estimated from Figure 2 of Repp (1980).

of the silence duration, and finds that there is no significant influence of variance on the adaptation process. Since the third experiment suggests that variance does not have an effect, the listeners' response must be based on the mean of the silent interval, within some time window. So, the final experiment tests the time interval in which the adaptation takes place.

3.2 Experiment 1: Replication of Repp (1980) cluster condition

In trying to replicate the results of Repp (1980) there were some difficulties during the pilot studies. Using synthetic speech tokens proved to be difficult since good exemplars of the CV /ib/ could not be created, and therefore, natural speech tokens were used. Thus, this experiment replicated the cluster condition of Repp (1980) using natural speech stimuli, in which the tokens were edited to closely approximate the Repp (1980) synthetic tokens.

3.2.1 Subjects

Six subjects participated in this experiment, including four graduate student volunteers, the author, and his advisor. An additional subject participated but was removed from the analysis due to his poor responses: he had no boundary between 1 and 2 stop consonants for the different anchor conditions. In addition, one student had knowledge of the Repp (1980) experiment, while the other students were naive subjects.

3.2.2 Stimuli

In order to get the VC and CV tokens, a male speaker uttered the syllable /ib-ga/ into a Sennheiser MD 421 microphone in a quiet room. The utterance was sampled

at 16 kHz using the Ariel dsp32c DSP board and stored on a Sun Sparcstation IPX. These tokens were cut into a VC (/ib/) and a CV (/ga/) token, removing the burst information and leaving only the voiced formant transition information. Thus, they resembled the synthetic stimuli in Repp (1980). In order to obtain only the voiced portions of the tokens, the signal was cut at the nearest zero crossing based on the time waveform and a spectrogram [2]. Listening tests of the CV and VC tokens verified that they were good tokens. Once the CV and VC tokens were obtained, the 11 different silent intervals were created at a sampling rate of 16 kHz. The 11 intervals ranged from 15 to 115ms in 10ms steps. Next the VC-silence duration-CV cluster tokens were created by appending the VC token, an appropriate silence duration, and the CV token.

All of the stimuli for each of the anchor conditions was saved in a separate file on the Sun Sparcstation. At the beginning of each condition, there were 20 examples of the cluster tokens, taken from the extremes of the silent intervals for that particular condition, presented in alternation. Following these examples, there was a 10 second gap, followed by the actual experiment. The actual experiment consisted of 99 tokens for the no anchor condition, and 100 tokens for the anchored conditions (Figure 3·3), with a 2.5 second inter-token interval. For the no anchor condition, there were 9 tokens of each of the 11 silence intervals. In the low anchor condition, there were 30 tokens of the 15ms silence duration, and 10 tokens each of the 25 to 85ms silence durations. The high anchor condition had 30 tokens of the 115ms silence duration, and 10 tokens each of the 45 to 105ms silence durations. The mean and standard deviation of the silent interval distributions were 43ms and 24.94ms for the low anchor, 65ms and 31.78ms for the no anchor, and 87ms and 24.94ms for the high anchor condition. The order of the tokens were randomized for each condition.

[2]The Entropic Research Laboratory's ESPS software package was used for manipulating the speech tokens.

3.2.3 Procedure

The subjects were presented with six tokens to familiarize the subjects with the tokens and to determine a comfortable listening level. The six tokens consisted of alternating tokens taken from the extremes of the closure durations (15ms and 115ms). The experiment consisted of one session with a break between each of the three conditions. Subjects were seated in a quiet room, and listened to the stimuli over AKG K130 headphones at a comfortable level. The subjects were prompted by the computer to initiate each of the anchor conditions, only being allowed to stop at a break. The presentation order of the anchor conditions was balanced across the six subjects. The subjects responded by circling "g" or "bg" on the response sheet corresponding to the number of stop consonants they heard. The subjects were also told that they should not expect equal number of responses between the two alternatives "bg" and "g" for the different conditions.

3.2.4 Results and discussion

The results for the six subjects are shown in Figure 3·5, and the pooled results are shown in Figure 3·6. The figures show the average number of stop consonants heard as a function of the distribution. Thus, hearing one stop corresponds to hearing only "g", whereas hearing two stops corresponds to hearing "bg." Figure 3·5 shows that the individual subject data is quite variable and "noisy" across subjects. However, the pooled results is less variable, and shows a pronounced shift for the three conditions. The 50 % point of the pooled responses are close to the means of the three anchor conditions. Statistical analysis using ANOVA shows that the three curves corresponding to the different anchor conditions are significant ($p < 0.0001$), and the silent interval values were significant ($p < 0.0001$). The slope for the no anchor condition in the pooled results is more shallow than the low and high anchor

conditions.

The results from this experiment reinforce the results of Repp (1980) showing that the psychometric curves shift based on the distribution of the silent intervals, i.e. the range and frequency (number of tokens) of the tokens. In addition, the slope change in Repp (1980) is replicated in these results for the no anchor condition across subjects. However, due to the variability of the psychometric curves across subjects, the slope change could be due to pooling across subjects, and not due to any true adaptation process. In other words, since each individual's psychometric curve for the no anchor condition have drastically different means, by averaging across the listeners, the pooled response curve becomes more shallow.

While these results support the results of Repp (1980), five out of the six subjects reported hearing /id-ga/ or /ida/ for quite a few tokens. This could be due to interaction between the VC and the silent interval, i.e. there could be cue trading issues between the initial formant transition and the silent interval. Another possibility is that this percept could be due to an alveolar flap percept [3] promoted by a very short closure duration. Dorman and Raphael (1980) showed that listeners perceive a stop consonant, whose formant transition correspond to /b/ or /g/, as /d/ for very small closure durations. In addition, the final formant frequencies for /b/ are not good cues in front vowel context (/i/). Due to the /d/ percepts and the questionable nature of the slope change, the second experiment replicated this experiment using the stop consonant /d/ in a back vowel context (/a/), which has better formant transition cues.

[3]An alveolar flap corresponds to the production of the stops /t/ or /d/ in which the tongue rapidly and very briefly produces a closure.

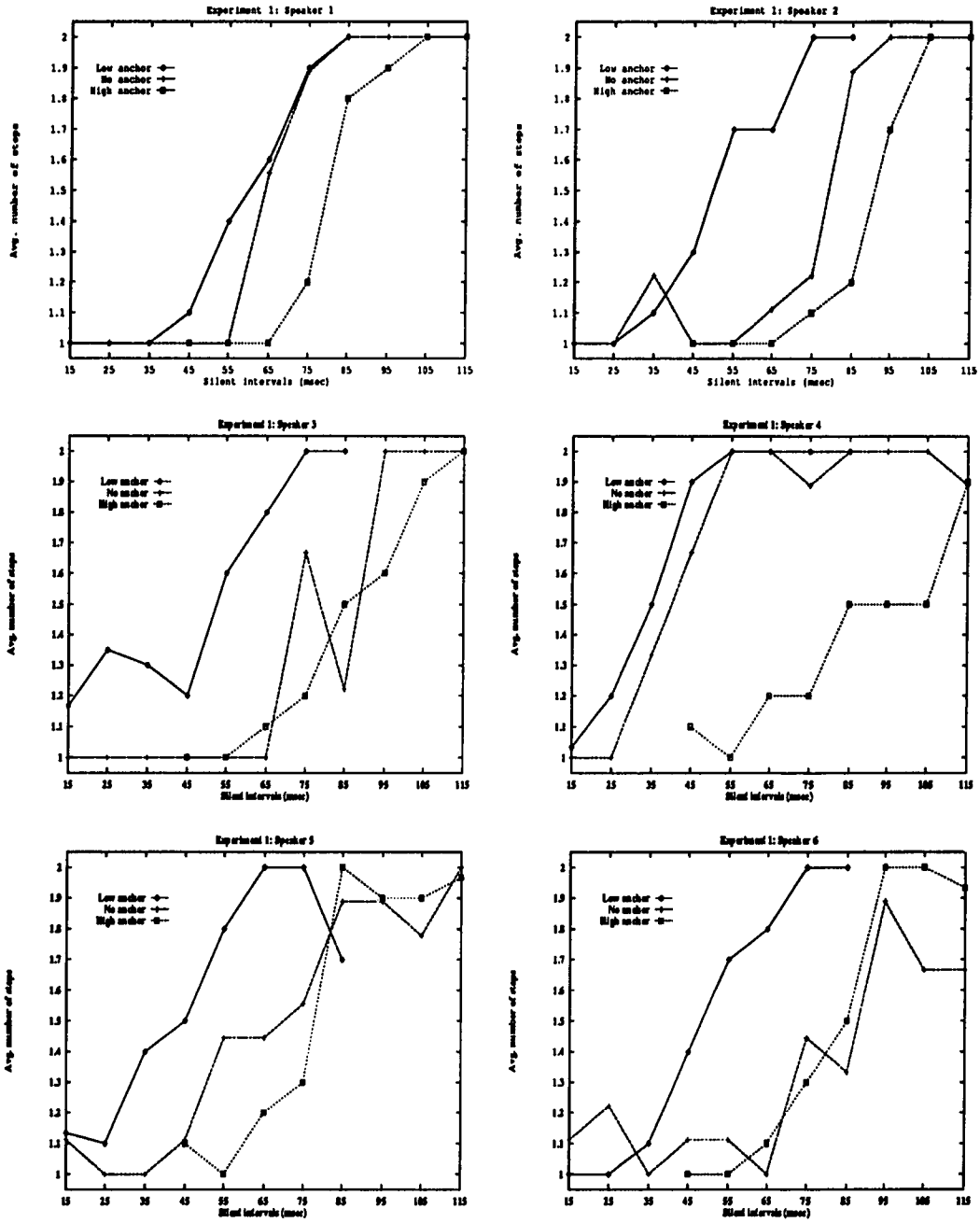


Figure 3·5: Results for experiment 1 for the six subjects for the low, no, and high anchor conditions. The figures show the average number of stop consonants heard for each silent interval.

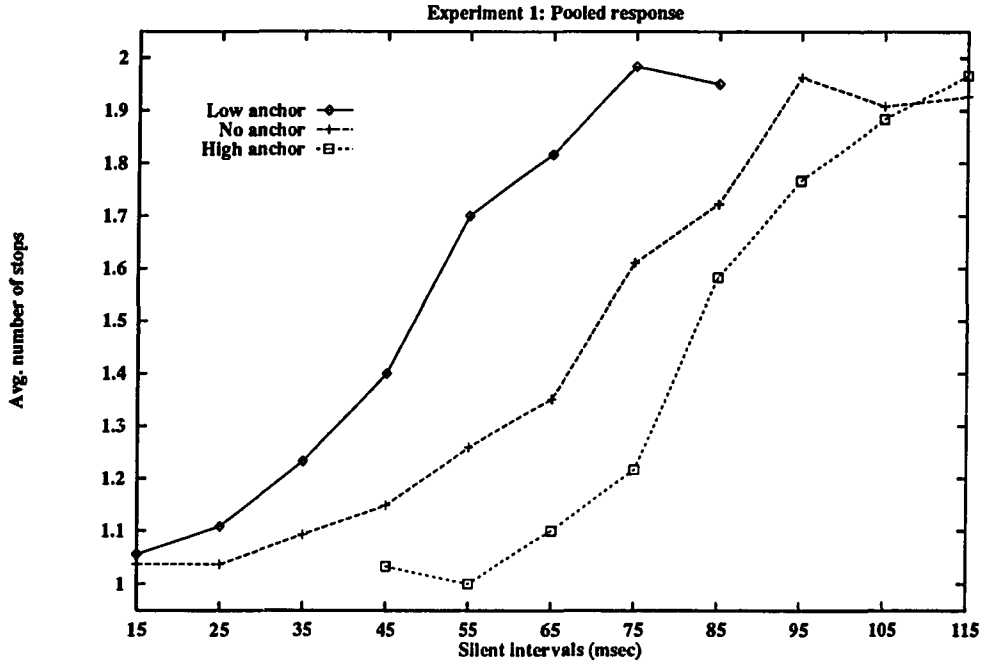


Figure 3·6: Results from experiment 1 pooled across the subjects for the low, no, and high conditions.

3.3 Experiment 2: Different place of articulation

This experiment tested the response of listeners in identifying the number of stops with a different VC, /ad/. Thus, this experiment used /ad-ga/ instead of /ib-ga/ tokens.

3.3.1 Subjects

The six subjects that participated in the first experiment took part in this experiment as well.

3.3.2 Stimuli

The stimuli were created in the same manner as experiment 1, except the original utterance was /adga/. The same male speaker uttered the syllable /adga/, which was broken up into burstless VC (/ad/) and a CV (/ga/) token. Using these VC and CV tokens, the exact method of obtaining the three files stated in experiment 1 was followed. The same randomization procedure was followed, and so the order of the tokens remained the same as in experiment 1.

3.3.3 Procedure

The procedure for this experiment was the same as experiment 1 except that the subjects responded by circling "d" or "dg" on the response sheet. Once again, the order of the conditions were randomized across subjects.

3.3.4 Results and discussion

The results for all six subjects are shown in Figure 3·7, and the pooled responses are shown in Figure 3·8, respectively. The individual subject data is less variable compared to the individual subject data from Experiment 1 (Figure 3·5). In addition,

the individuals show shifts of the psychometric curves. The subjects found this experiment much easier to perform than the previous experiment, which is reflected in the subjects' performance. Once again, the pooled result shows a shift in the curves for the three different conditions, with the 50% points being close to the means. The analysis of variance shows that the three anchor conditions are significantly different ($p < 0.0001$), as well as the durations themselves ($p < 0.0001$).

The pooled results do not show any apparent slope change across the three conditions, implying that the slope change seen in Repp (1980) and in Experiment 1 are probably due to poor formant transition cues specifying /b/, or due to the distractions from flap percepts, or due to pooling across subjects.

Since Experiment 1 and 2 differed only in the initial VC, ANOVA tests were performed across experiments for each of the different anchor conditions to see if cue trading exists. The results from ANOVA show that the no and high anchor conditions differed significantly between Experiment 1 and 2 ($p < .04$), but the low anchor condition was not significantly different ($p < .33$). Thus, the result seems to implicate some cue trading; however, in light of the individual variability and the flap percepts in Experiment 1, the significance of the differences between Experiment 1 and 2 is questionable and needs further exploration.

Given that the listener is adapting to these different distributions, the listener could normalize the silent interval distribution based on the mean and/or variance. Therefore, the next experiment tests whether the variance of the overall distribution has any influence. Also, since the /ad/ token was much easier to perceive than the /ib/ token, it was decided to use the /ad-ga/ tokens for the rest of the experiments.

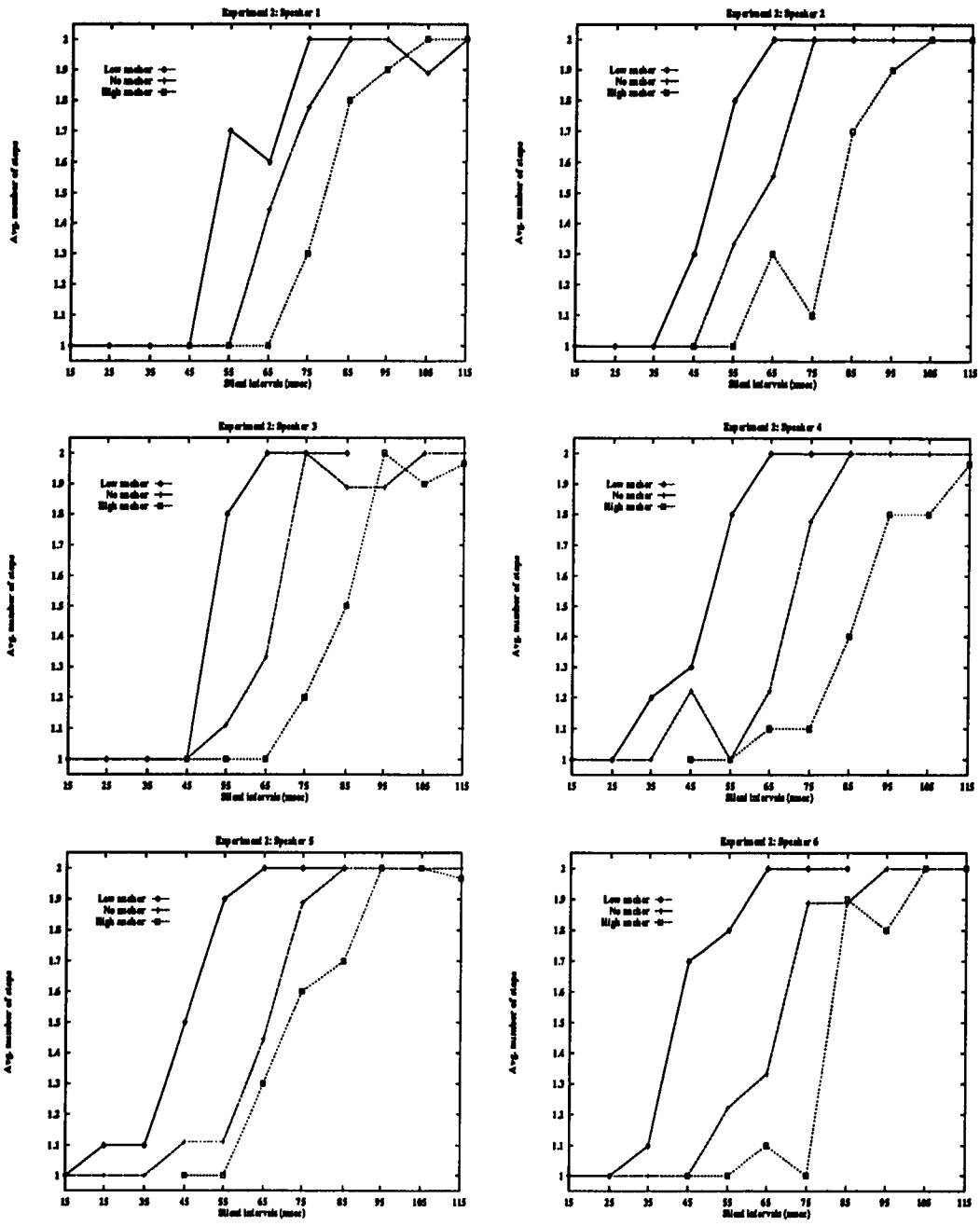


Figure 3-7: Results for experiment 2 for the six subjects for the three anchor conditions. The figures show the average number of stop consonants heard for each silent interval.

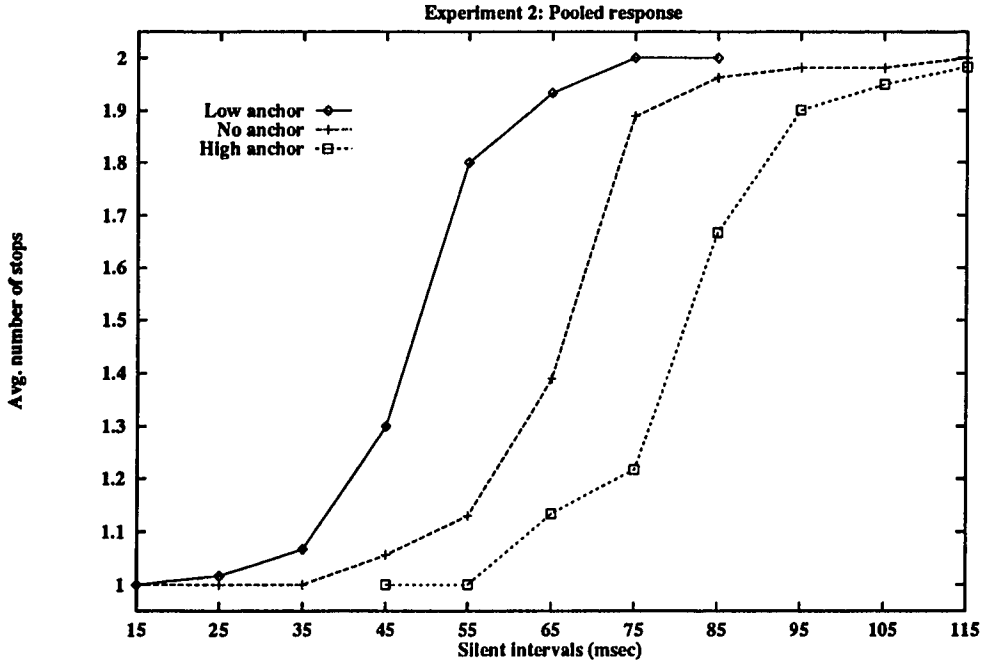


Figure 3·8: Results from experiment 2 pooled across the subjects for the three conditions.

3.4 Experiment 3: Effects of silent interval distribution variance

This experiment examines the effects of the overall variance of the silent interval distribution. In order to test this, three different silent interval distributions were chosen. The distributions kept the mean the same as the no anchor condition in experiment 2 (65ms), but varied the overall variance of the distribution. The first condition is a replication of the no anchor condition from experiment 2, denoted here as the normal range condition. The small range condition consists of the normal range condition with only the center 9 intervals, and thus, a smaller variance (25.95ms) than the normal range condition (31.78ms). The U anchor condition has a larger variance (36.70ms) than the normal range condition. This is accomplished by having more tokens of the smallest and largest intervals than the other intervals. The distributions are shown in Figure 3·9. A chi-square test was performed. It verified that the three distributions do significantly differ ($p < .025$).

3.4.1 Subjects

The six subjects that participated in the first and second experiment took part in this experiment as well.

3.4.2 Stimuli

The stimuli that was used consisted of the same /ad/ and /ga/ tokens as in Experiment 2. However, the distributions of the silent intervals were changed (Figure 3·9). There were three different conditions: no anchor or normal range, small range, and U anchor. The normal range condition consisted of the same silent interval distribution from the no anchor condition from the second experiment. In the small range condition, the range of the silent intervals went from 25 to 105ms with 11 tokens of each

stimuli. In the U anchor condition, the smallest (15ms) and largest (115ms) silent interval had 18 tokens each, with the 7 tokens each for the intervals 25 to 105ms. So, all three conditions had 99 tokens. With these distributions, the cluster tokens and three files, corresponding to the three conditions, were created in the same manner as the no anchor condition in experiment 2. The tokens were once again randomized.

3.4.3 Procedure

The experiment consisted of one session with two breaks between the three conditions. Once again, the order of the anchor conditions was balanced across all the subjects. Other than this, the procedure was the same as in experiment 2.

3.4.4 Results and discussion

The results for each subject is shown in Figure 3·10, and the pooled results are shown in Figure 3·11. For most of the subjects, the curves corresponding to the three conditions are quite similar, and there is little variability across subjects. In fact, variance analysis found that the three different conditions did not differ significantly ($p < 0.4203$). The pooled results are similar in that there is little difference across the three conditions.

Thus, the results suggest that the effect of overall variance is insignificant, and that the listener must be using only the mean to adapt to the silent interval. The fourth experiment manipulated the temporal distribution of the silence intervals to gain insight into the "window" over which the listeners compute the average silence duration.

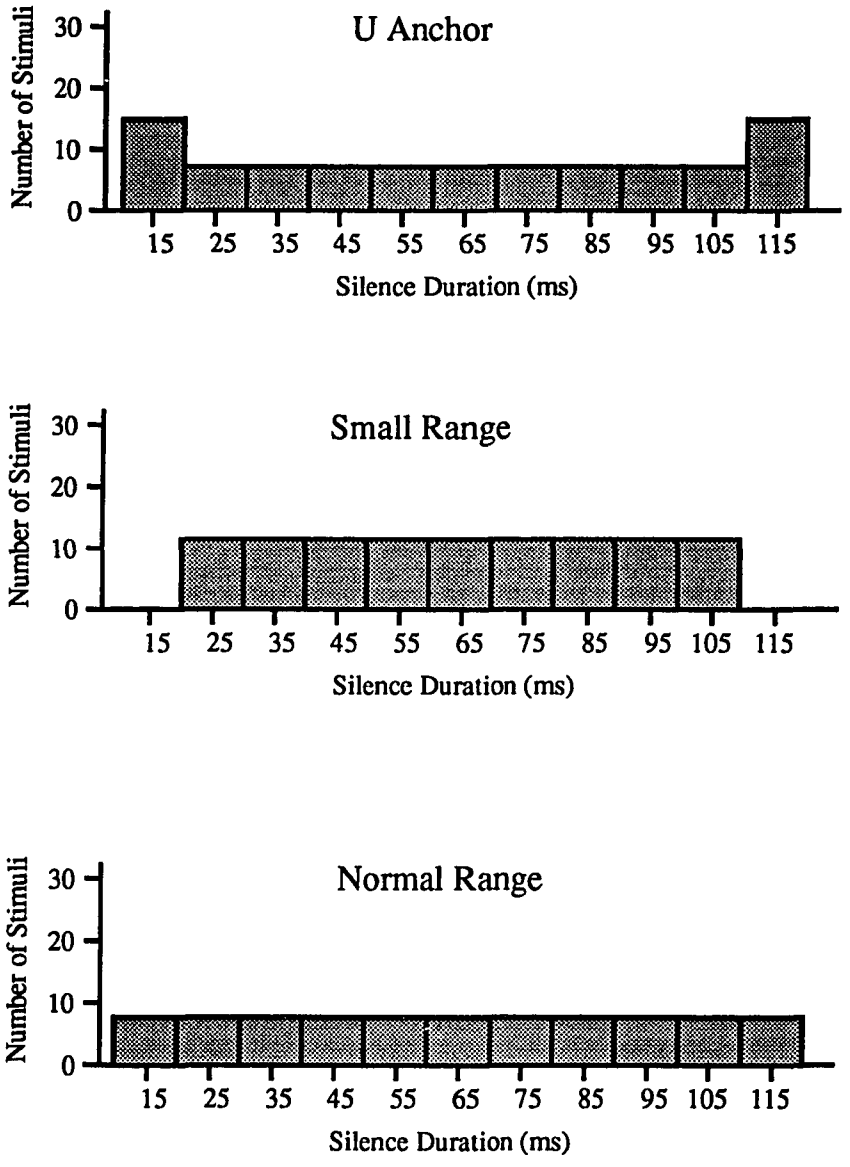


Figure 3-9: Distribution of the silent intervals for the small range condition, the U anchor condition, and the normal range condition.

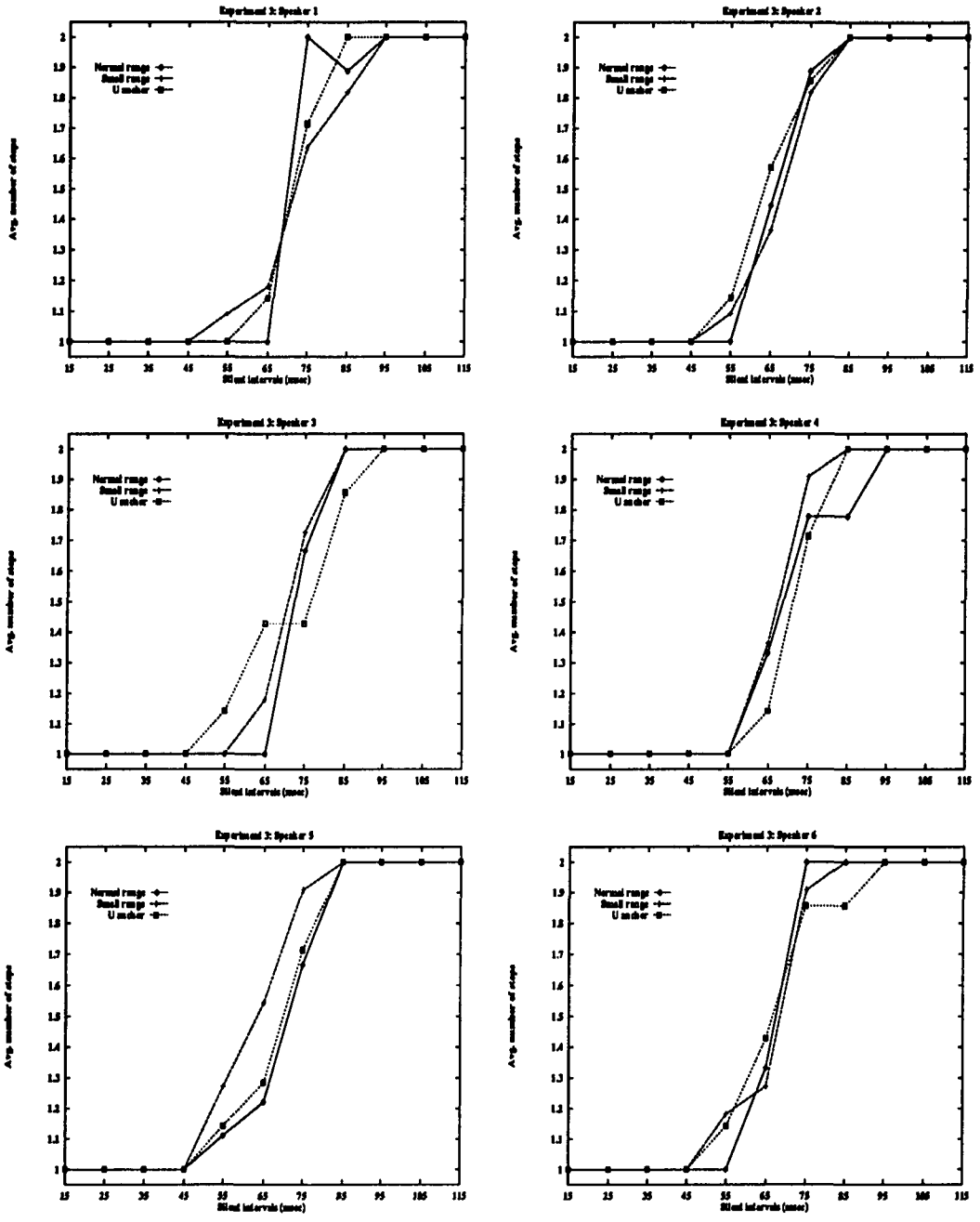


Figure 3·10: Results for experiment 3 for the six subjects for the small range, U anchor and the normal range. The figures show the average number of stop consonants for each silence duration.

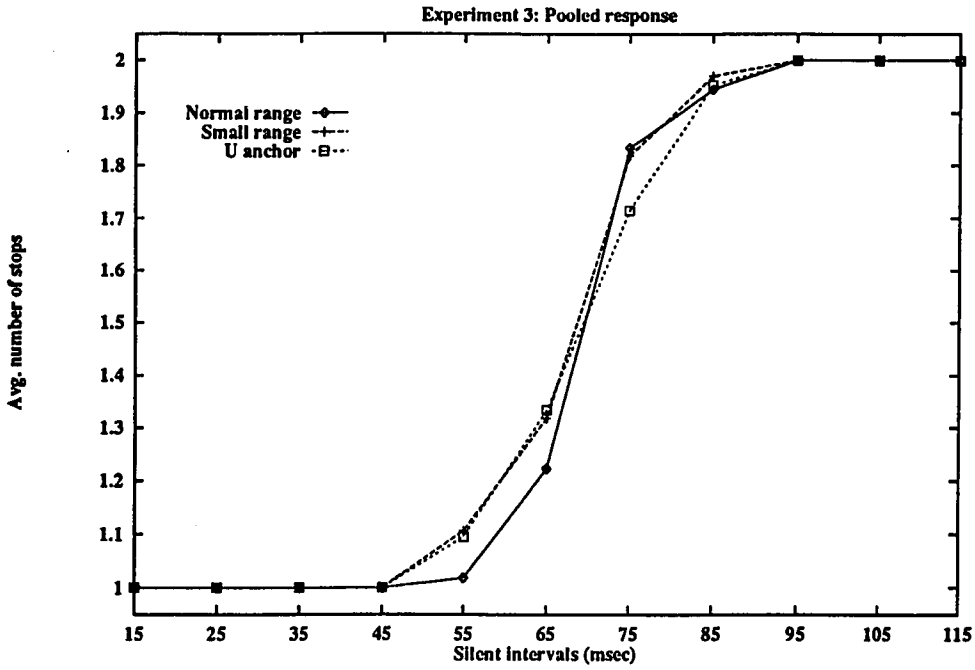


Figure 3-11: Results from experiment 3 pooled across the subjects for the three conditions.

3.5 Experiment 4: Temporal characteristics of adaptation

Since the previous experiments suggest that listeners are adapting to the mean silent interval, this experiment investigates the influence of the temporal token presentation on the stop consonant cluster percept by testing the prior two tokens' influence on the percept of the current token. Listeners are presented with M trials, where each trial contains 4 different tokens of /ad-ga/: P, P', T, and T'. P and P' are the two prior tokens, and the subject's response for the test token, T, is obtained. The M trials correspond to the different values of the test token T, which are used to determine psychometric curves.

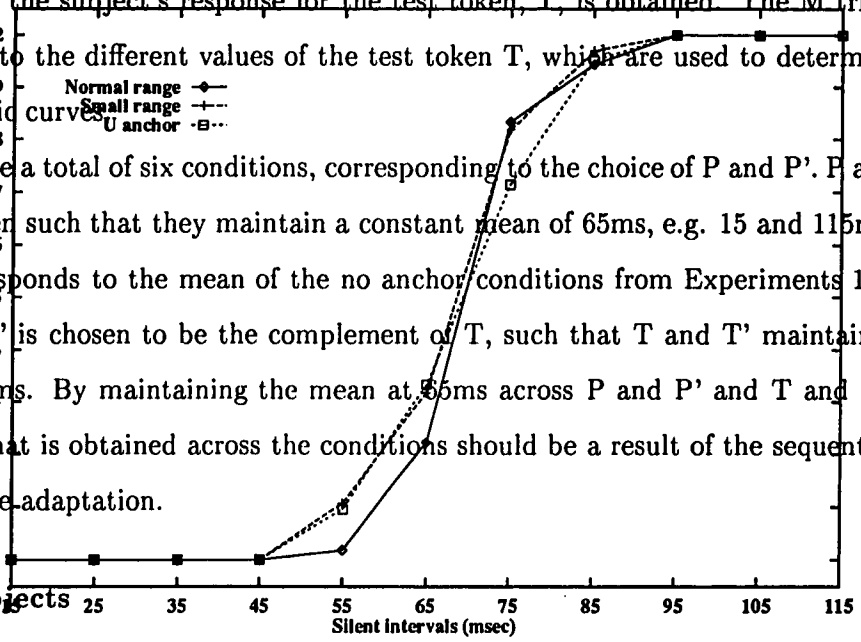


Figure 3-11: Results from experiment 3 pooled across the subjects for the three conditions.

There are a total of six conditions, corresponding to the choice of P and P'. P and P' are chosen such that they maintain a constant mean of 65ms, e.g. 15 and 115ms, which corresponds to the mean of the no anchor conditions from Experiments 1-3. Similarly, T' is chosen to be the complement of T, such that T and T' maintain a mean of 65ms. By maintaining the mean at 65ms across P and P' and T and T', any effect that is obtained across the conditions should be a result of the sequential nature of the adaptation.

3.5.1 Subjects

Five of the six subjects that participated in Experiments 1, 2, and 3 took part in this experiment.

3.5.2 Stimuli

The stimuli consisted of a subset of the stop consonant cluster tokens (/ad-ga/) that were used in experiments 2 and 3. The /ad-ga/ tokens consisted of six cluster tokens with the following silence duration: 15ms, 35ms, 55ms, 75ms, 95ms, 115ms.

There were six conditions corresponding to the different P and P' presentations:

15 and 115ms, 35 and 95ms, 55 and 75ms, 75 and 55ms, 95 and 35ms, and 115 and 15ms. Each condition contained a total of 30 trials. Each trial consisted of four tokens: P, P', T, and T', where P and P' are set to the values dictated by the particular condition. The 30 trials correspond to five repetitions of the six different cluster test tokens, T. T' is chosen to be the complement of T, such that T and T' maintain the mean at 65ms. For example, if T is 35ms, then T' is chosen to be 95ms. The order of the test tokens T was randomized for each condition. All the tokens for each condition was placed in a separate file on a Sun Sparcstation.

Since the inter-token interval and the inter-trial interval was 1 second, the subjects had no knowledge that there were multiple trials in each condition.

3.5.3 Procedure

Each condition had six alternating examples from the extremes of the condition, followed by a 10 second gap, followed by the actual condition. The experiment consisted of one session with breaks between the six conditions. The order of presentation of the different conditions were randomized across speakers. The subjects responded by circling "d" or "dg" on the response sheet for every token. Thus, the subjects were unaware of the differences between the tokens. Only the response for the test token T was used in determining the psychometric curves.

3.5.4 Results and discussion

The results for each subject is shown in Figure 3·12, and the pooled results are shown in Figure 3·13. For most of the subjects, the curves corresponding to the different conditions are quite similar, but slightly different. The six conditions do differ significantly ($p < .02$), and comparison of the different P-P' pairs show significance. 15-115ms and 115-15ms ($p < .04$), 35-95ms and 95-35ms ($p < .03$), and 55-75ms and

75-55ms ($p < .04$).

This experiment shows that there is a sequential component to the adaptation process since the P-P' pairs, e.g. 15-115ms and 115-15ms, are significantly different. Since the mean was kept the same, the resulting differences have to be due to the order of presentation. Furthermore, an interesting result that is seen across most subjects and in the pooled response is the reversal of the curves for (P-P') for 115-15ms and 15-115ms. In other words, the psychometric curve for 115-15ms is to the right of the 15-115ms curve, which is the opposite of what one expects if the previous token has the most weight in determining the location of the psychometric curve. This could be due to contrast effects or a smaller weighting to the prior token.

However, the experiment contains some possible problems. If the adaptation window is not flat and is greater than three tokens, then there is the possibility of interaction between prior trials and the current trial. Thus, this experiment needs to be extended to determine the weighting and the window size.

3.6 Linear adaptation model

Based on the results of Repp (1980), Boardman, Cohen, and Grossberg (1993) created a model which emulated the percept of the stop consonant clusters. In the model, identification of the second stop consonant was based on an adaptation process that found the mean silent interval. The model was able to replicate the psychometric curves of Repp (1980), although the slope change in the no anchor condition was not obtained. In order for the model to produce the slope change for the no anchor condition a bias had to be introduced for each "subject." Thus, the slope change is obtained by averaging across these "subjects." In other words, the model predicted that the slope change was due to averaging across subjects, which was validated in the results of these experiments.

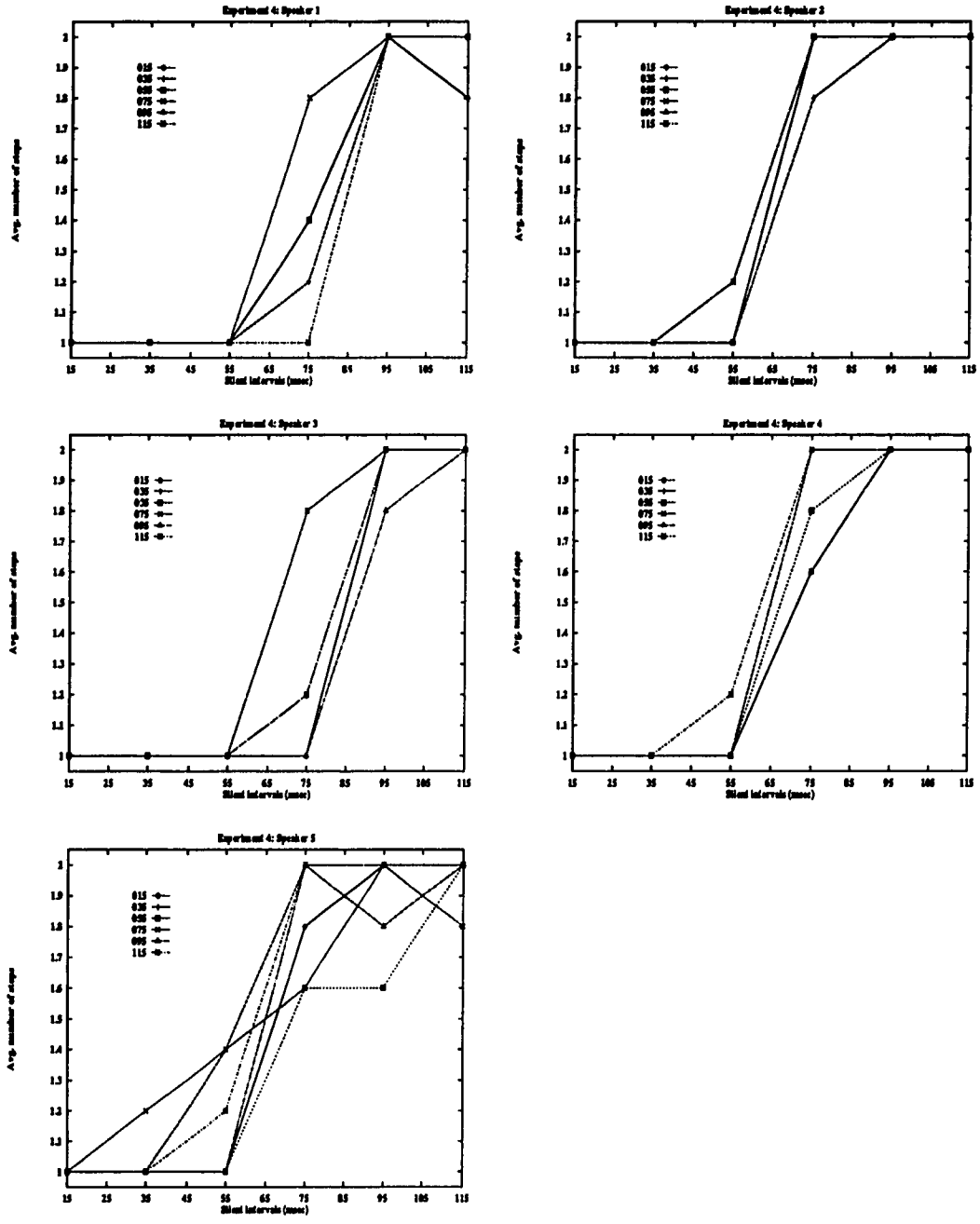


Figure 3·12: Results for experiment 4 for the five subjects for the different conditions. The curves correspond to the different P' conditions, as listed.

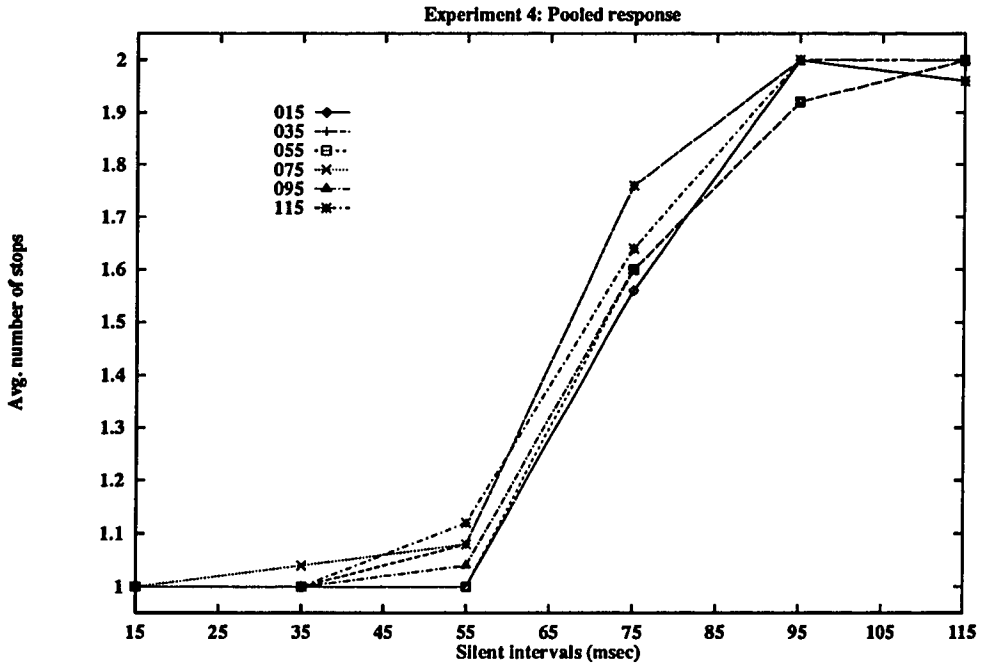


Figure 3·13: Results from experiment 4 pooled across the subjects for the different conditions. The curves correspond to the different P' conditions, as listed.

In order to study the adaptation process in more depth, a simple model of the adaptation process was analyzed. This adaptation model was used to fit the psychometric functions and then used to predict the subjects' performance on other conditions. Moreover, since it was determined from experiment 3 that listeners do not use the variance of the distribution to adapt to the silent interval, the simplest assumption is that the listener adapts to the mean silent interval within some time window.

The model assumes that there is a decision boundary between hearing 1 and 2 stop consonants which shifts as a function of the prior silent intervals. The model derives the mean silent interval within a specified time window. The window could correspond to an exponential weighted average or a flat running average. After the mean is found, the model subtracts this weighted average from the current silent interval to derive a decision. Furthermore, it is assumed that there is white (gaussian) noise, with a zero mean and unit standard deviation, which perturbs the judgement process. The model is described as follows:

$$N(t) = H(S(t) - \sum_{k=1}^{M} a_k S(t-k) + \epsilon(t)) + 1, \tag{3.1}$$

where N(t) is the number of stops heard by the model at time t, $S(t)$ is the value of the silent interval at time t, a_k are weighting factors less than 1, M corresponds to the window length, $\epsilon(t)$ is white noise corresponding to noise in the perception process, 1 represents the fact that the second stop is always heard, and $H(x)$ is a function, denoting the all-or-none percept of the first stop consonant:

$$H(x) = \begin{cases} 1 & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Equation 3.1 states if the current silent interval, $S(t)$, is greater than both the

mean silent interval, $\sum_{k=1}^{M} a_k S(t-k)$, and the spurious noise, $\epsilon(t)$, then one hears the first stop, and a total of 2 stops. However, if the current silent interval $S(t)$ is less than the mean and the noise, $H(\ldots) = 0$, and so, only the second stop is heard. The effect of the noise is to produce a 50% probability of hearing the first stop when the current silent interval is at the mean silent interval.

The weighting is derived from a particular subject's responses by using the perceptron learning rule. The perceptron learning rule states that if the output of the network equals the target value, then no change in the weights are made. However, if there is an error, then a weight is changed in proportion to the error multiplied by the input. During the training phase the noise $\epsilon(t)$ is set to 0.

Since the first experiment proved to be more difficult for the subjects, that experiment was not modeled. For experiment 2, the weights were obtained by training the network using the no anchor condition. After training, the network's performance was determined on the no anchor condition as well as the novel tokens from the low anchor and the high anchor conditions. This was done for experiment 3 as well: the model was trained on the normal range condition, and then tested on the normal range condition as well as the novel tokens from the U anchor and small range conditions. In deriving the training data, the subjects' response for the first M tokens were based on the example tokens, which alternated between the extremes of the condition.

3.6.1 Window length and weighting

One method for gaining insight into the window length is to see how the error changes as the window length M is varied. Figure 3·14 and 3·15 shows the resulting errors as the window length is varied from 1 to 15 tokens for each subject for Experiment 2 and 3, respectively.

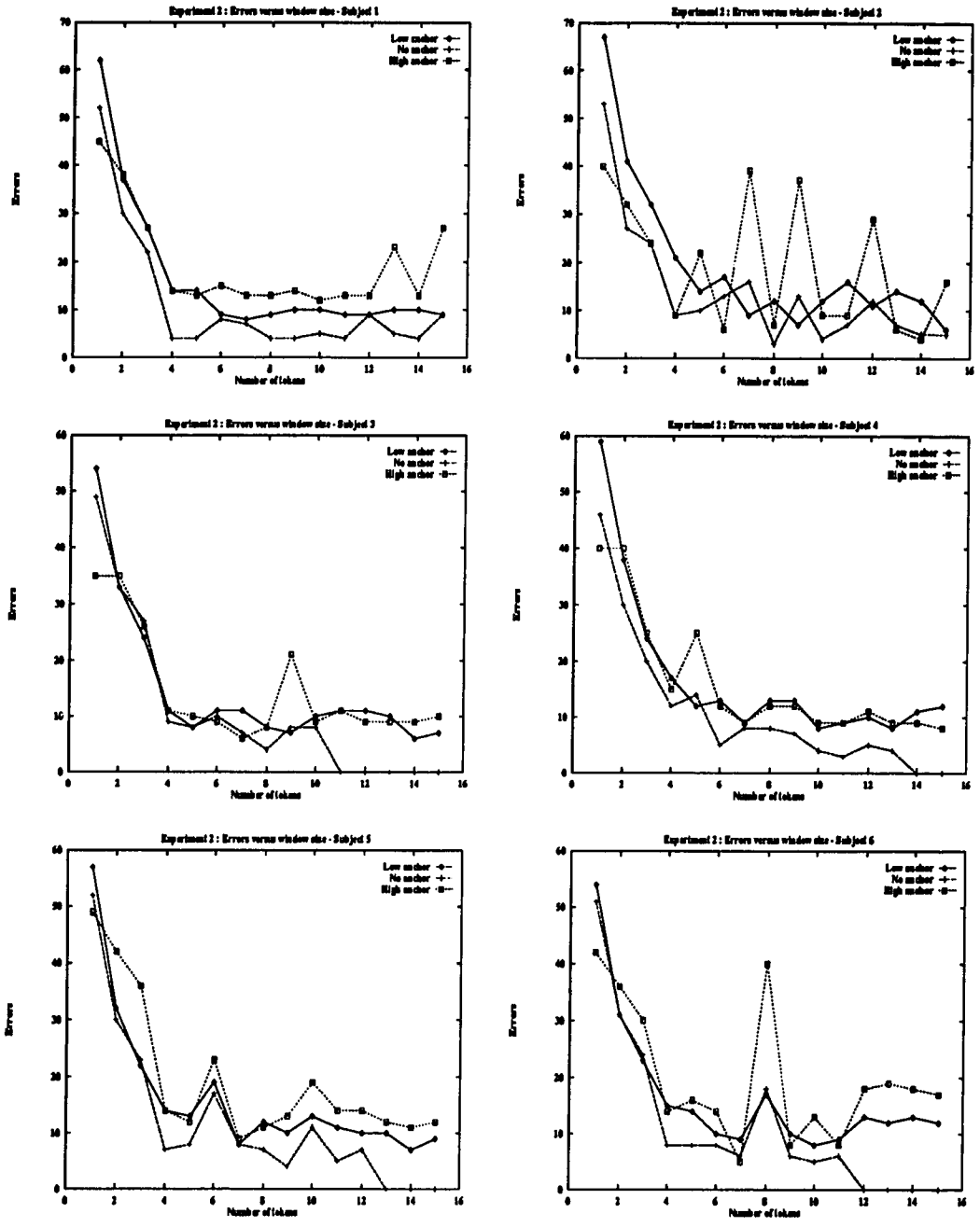


Figure 3-14: Errors as a function of window size for each subject's data from experiment 2.

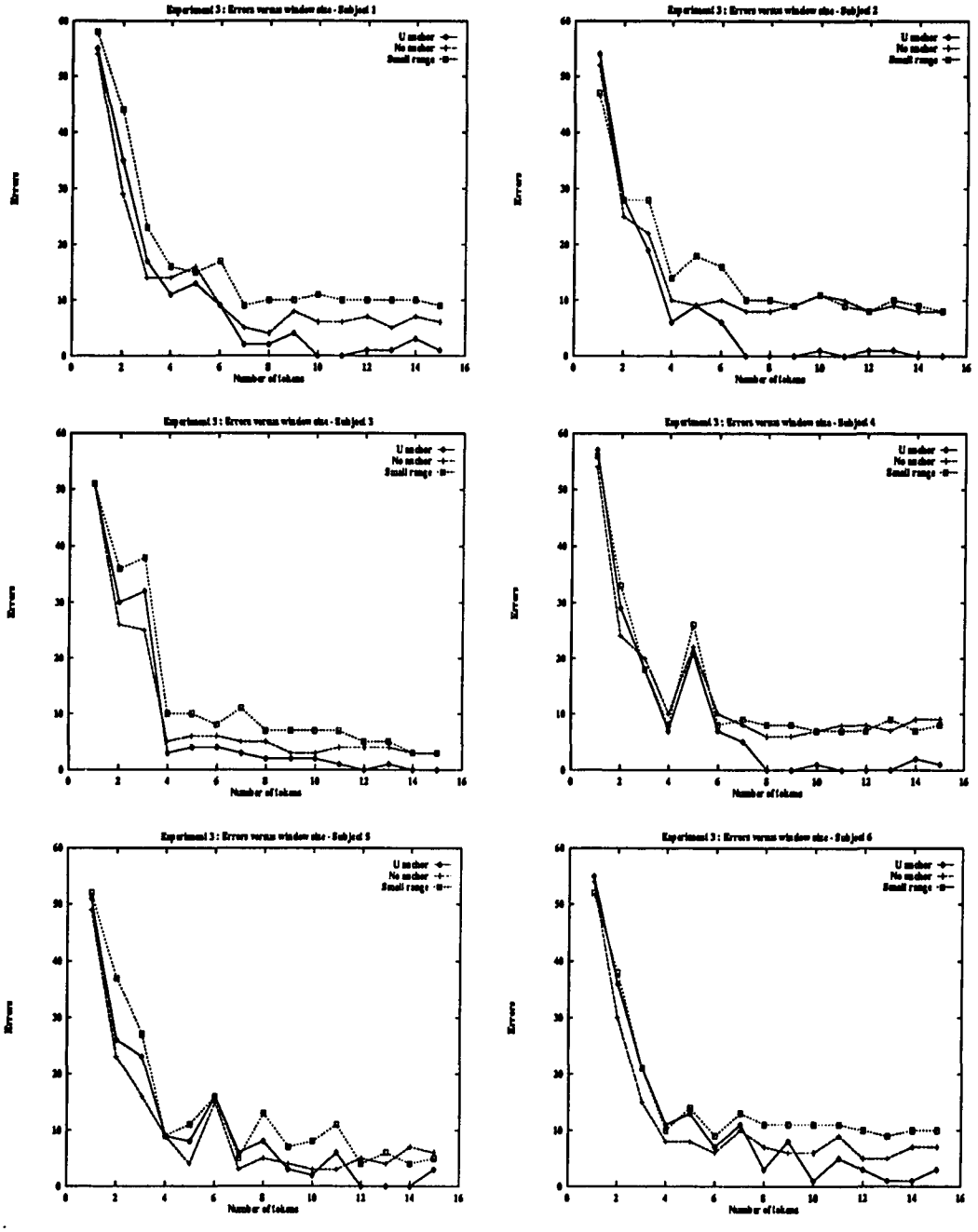


Figure 3·15: Errors as a function of window size for each subject's data from experiment 3.

For the most part in both Experiment 2 and 3, the errors tend to plateau after 7. Thus, the implication is that the window size within which the listener adapts is less than 6 tokens, and that more tokens from the past do not provide more information. The weights a_k are shown for a window length of 7 tokens in Figures 3·16 and 3·17 for Experiments 2 and 3, respectively. The weights seem quite chaotic across subjects. This could be due to the fact that the linear model has too many parameters a_k which dictate the result, and thus, more variability in the weights. A better model might consist of an exponentially weighted average, in which there are fewer parameters.

3.7 General results and conclusion

The results from these experiments support the results of Repp (1980), which showed that there are range and frequency (number of tokens) effects of silence duration in stop consonant cluster perception. The first experiment showed that the psychometric curves, corresponding to the number of stop consonants heard in an /ib-ga/ context, shifted and changed slope as a function of the silent interval distribution shape. However, the individual subject data was quite "noisy" and variable, and subjects heard /ida/ and /id-ga/ percepts for some of the tokens. These anamolies are probably due to flap percepts and poor formant transition cues signifying /b/ due to the front vowel context.

The second experiment showed that the psychometric curves exist for a different stop cluster /ad-ga/. In addition, the subject data is less variable, and no slope change is evident, lending credence to the notion that the slope change in the first experiment is due to averaging across noisy subject data.

The issue of cue trading between the formant transitions and the silent interval needs further exploration. While a comparison between Experiment 1 and 2 showed significant differences for two of the three conditions, the variable results of Experi-

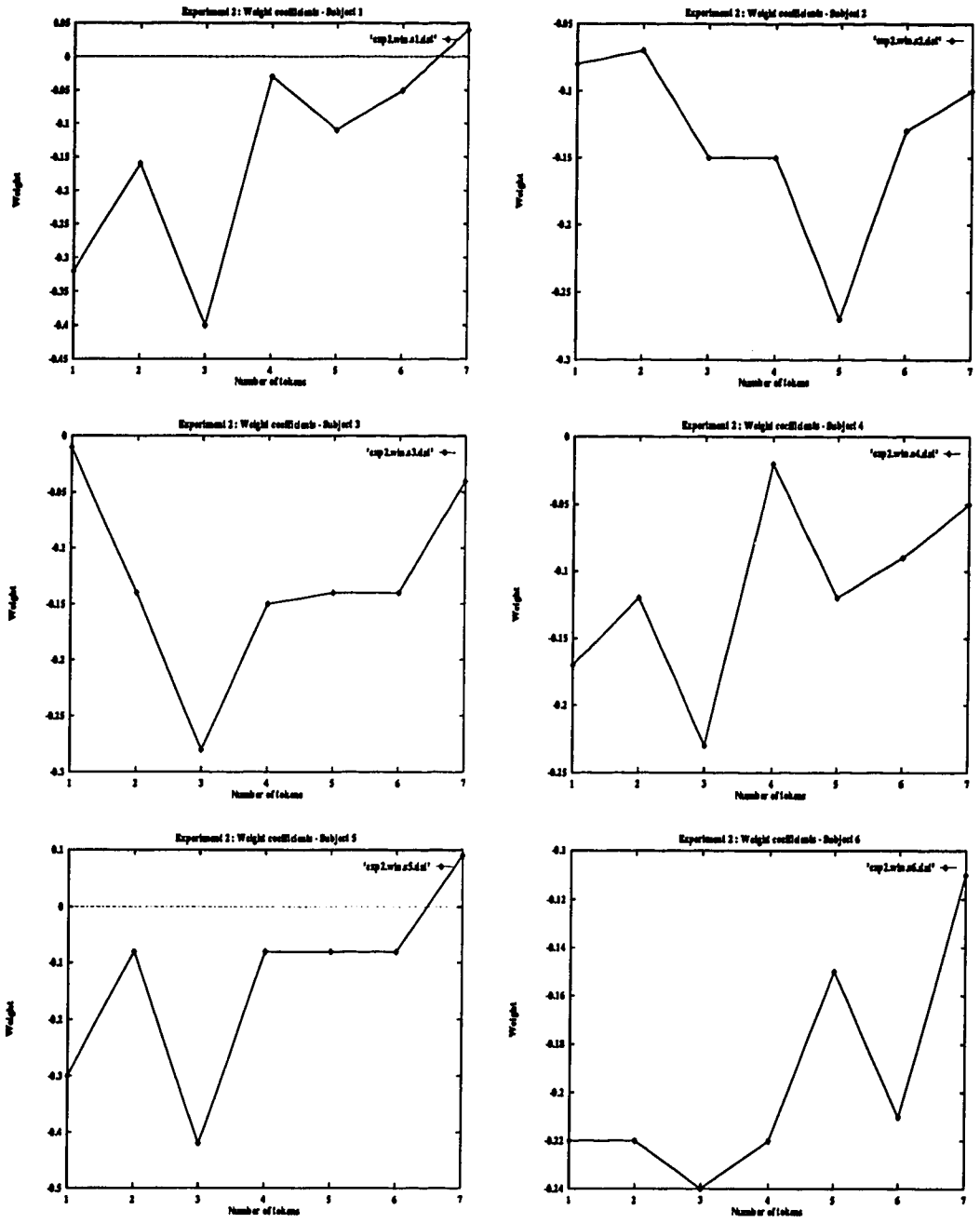


Figure 3-16: Weighting coefficients a_k for a seven token window length (M = 7) for each subject's data from experiment 2.

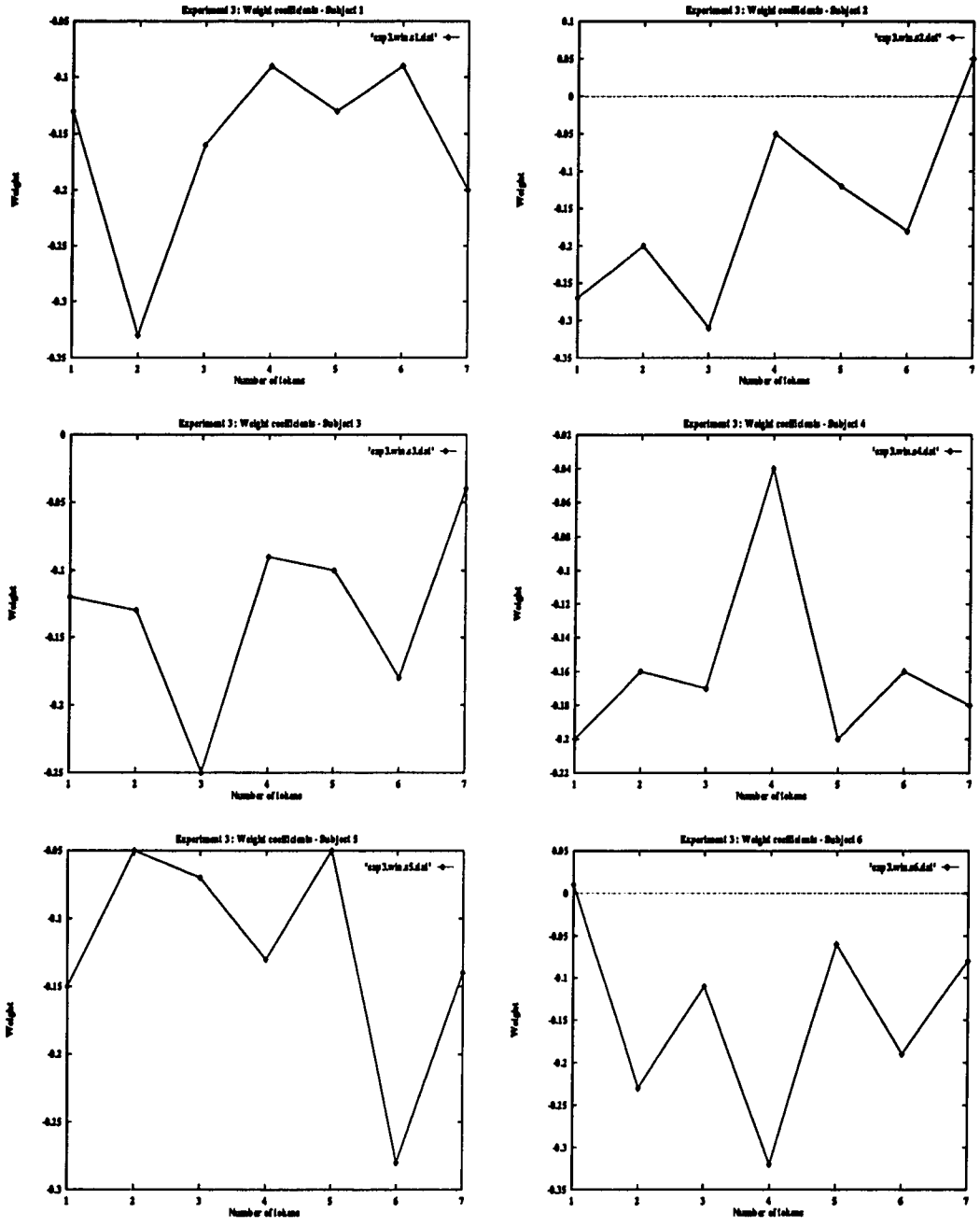


Figure 3·17: Weighting coefficients a_k for a seven token window length (M = 7) for each subject's data from experiment 3.

ment 1 confounds this issue. Thus, the issue of cue trading needs to studied further, and verify if there is a true interaction between formant transition cues and silence duration cues.

The third experiment supports the notion that while listeners are adapting to the distribution of the silent intervals, they do not employ the variance in this adaptation process. The fourth experiment shows that the adaptation process contains sequential effects, and that the window includes more items than just the prior token. While the results show there is some adaptation taking place, better data on the sequential aspect of this process is needed.

The linear model of adaptation provided insight into the length of the window by denoting when the error on the test tokens plateaued, which corresponded to 7 tokens. Since the modeling effort was not successful in providing insight into the shape of the window due to the number of weights, a model with fewer degrees of freedom should be explored.

Chapter 4

A neural network model of auditory scene analysis

4.1 Introduction

The ability of a listener to pay attention to a particular speaker in a noisy room or in a room with other speakers, e.g. at a cocktail party, attests to the robustness of the auditory perceptual system. Even though these multiple sound sources mix together their harmonics to produce one signal at the listener's ear, the auditory system is capable of teasing apart this jumbled signal to recognize different mental objects for the different sound sources. The ability to segregate these different signals has been termed auditory scene analysis (Bregman, 1990). The scene analysis corresponds to the mechanisms by which the auditory system selectively groups certain acoustic features, while excluding others, to form internal representations of auditory objects.

An analysis of the mechanisms of auditory scene analysis is important for understanding how the human auditory perceptual system operates, as well as for technological applications. While speech recognition systems have improved greatly within the last decade, they are still prone to noise and interference from other speakers.

4.1.1 Auditory scene analysis

The nomenclature associated with auditory scene analysis contains several keywords: source, stream, grouping and stream segregation. The source is a physical, external

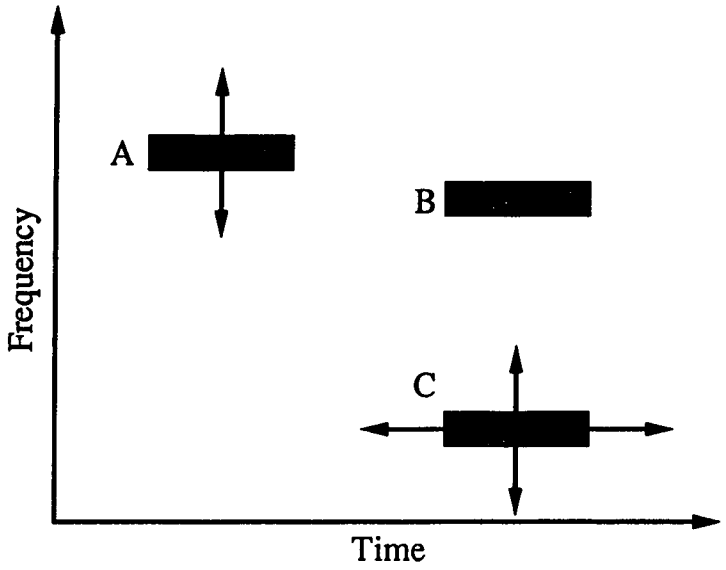


Figure 4·1: A groups better with B if they are closer in frequency. However, simultaneous cues, such as common onsets, common offsets and harmonicity, can help group B and C. After Bregman and Pinker (1978).

entity which produces sound; e.g. a speaker. The perceptual correlate of this source is a stream, i.e. it is "what the brain takes to be a single sound" (Bregman, 1984). The stream is created by the perceptual grouping and segregation of acoustic properties that are thought to correspond to an object. Grouping and stream segregation, or streaming, assign appropriate combinations of frequency components to a stream. For an exhaustive review of auditory scene analysis, the reader is referred to Bregman (1990).

The scene analysis process can be thought of as two processes that interact: a simultaneous grouping process and a sequential grouping process. For example, in Figure 4·1, the simultaneous grouping process tries to group B and C together if they have synchronous onsets and offsets, or if they are harmonically related. Similarly, the sequential grouping process tries to group A and B together based on their frequency and temporal proximity.

4.1.2 Grouping principles

In order to denote which acoustic attributes correspond to a stream, researchers, including Gestalt scientists and, more recently, Bregman (1990) and his colleagues, have suggested several grouping principles:

- Proximity

  The proximity grouping principle is shown in Figure 4·1. If two tones are closer together in frequency and time, then it is more likely that they should be grouped together, e.g. A and B should be grouped together if they are close enough.

- Closure and belongingness

  Closure and belongingness lead to percepts of continuity and completion. Closure is the perceptual phenomenon of completing streams when there is evidence for it. For example, listeners hear a tone continuing through noise (Figure 4·2), even though the tone is not present during the noise (Miller & Licklider, 1950). Thus, the perceptual system completes the tone across the noise, given the evidence that the same frequency tone is present on either side of the noise. This is also known as the auditory continuity illusion.

- Good continuation

  Good continuation states that an object's sound does not make rapid jumps, but instead continues smoothly. For example, in Figure 4·2 the slope of the tone is the same on either side of the noise, and thus should be grouped together due to good continuity of the tone. However, if the post-noise tone was at a distant frequency, then the tone would not have good continuity and would not stream across the noise. Note that continuity is closely related to proximity.

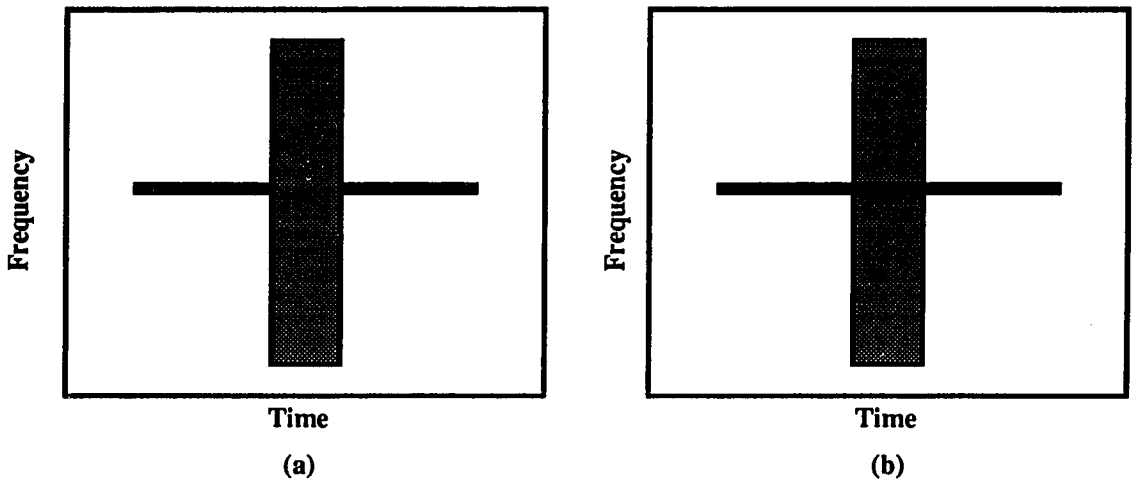


Figure 4·2: Stimulus and percept of the continuity illusion. (a) shows the stimulus that is presented to listeners, and (b) represents the percept. Note that in the stimulus, the tone does not continue through the noise, but stops at the onset of the noise, and continues at the offset of the noise, but the percept is that the tone continues through the noise.

- Common fate

  Common fate states that those attributes which are going through similar manifestations should be grouped together. For example, those frequency components which originate from the same spatial location share the same "fate", and therefore, should correspond to the same object. Similarly, those frequency components which are being modulated (frequency or amplitude) at the same rate or have synchronous onsets and offsets should correspond to an object.

- Principle of "exclusive allocation"

  This principle states that attributes are assigned to one stream or another, but not both. While this principle seems to hold in sequential streaming, it can fail in simultaneous streaming, where harmonics of two streams can overlap.

4.1.3 Primitive versus schema-based segregation

Bregman (1990) noted that auditory stream segregation consists of a primitive, non-attentive, unlearned process and a schema-based, attentive, learned process. Bregman and Rudnicky (1975) found that tones in an unattended stream can capture tones from an attended stream. In addition, van Noorden (1975) presented a repetition of two alternating tones whose frequency and temporal spacing were manipulated to subjects. van Noorden obtained two curves: the temporal coherence boundary (TCB) and the fission boundary (FB). The TCB corresponds to the boundary where the frequency separation between the temporally adjacent tones was too large to hear one stream. The FB corresponds to the point where the two frequencies were too close in frequency to be heard as separate streams. The FB varied little as a function of the tone repetition rate, and was mainly a function of the frequency separation. On the other hand, the TCB showed that as the frequency separation between the tones increased, one needed to slow down the repetition rate in order to maintain one stream with both tones. Bregman (1990) argued that the FB corresponds to an attentional mechanism and the TCB corresponds to non-attentional mechanism, and noted that the schema-based mechanisms can override the primitive mechanisms. The mechanism proposed here addresses the pre-attentive, primitive segregation mechanisms.

4.2 Grouping cues

One can find acoustic attributes that correspond to the grouping principles. The attributes include temporal and frequency separation, harmonicity, spatial location, amplitude modulation, frequency modulation, and onsets and offsets.

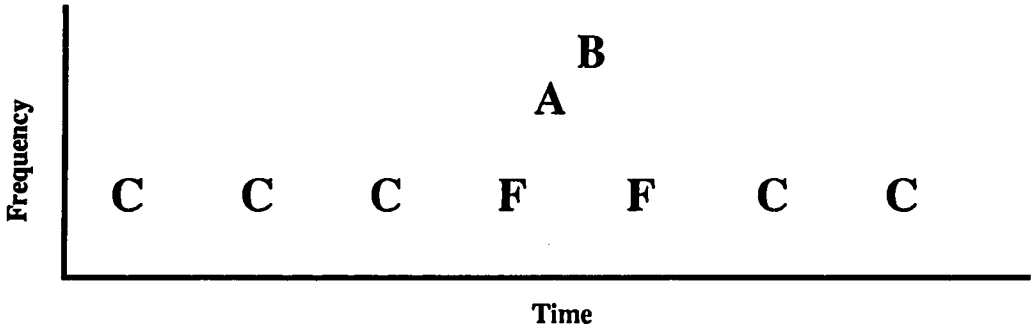


Figure 4·3: When A and B are presented by themselves, listeners could easily judge the order of them. If A and B were flanked by tones F, then listeners had a more difficult time. However, if the captor tones C surrounded the flankers, then F streamed with C, leaving A-B to a different stream, allowing the listeners to hear the order once again. After Bregman and Rudnicky (1975).

4.2.1 Temporal and frequency separation

Bregman and Pinker (1978) showed that tones in a repeating sequence tend to group if they are closer in frequency, e.g. A and B in Figure 4·1. In addition, faster presentation rates of alternating high and low frequency tones causes the two tones to be segregated into 2 streams (Bregman and Campbell, 1971). The effect of faster presentation rates is to narrow the temporal separation between adjacent instances of the high tone (and low tone), allowing for streaming of the high tone (and low tone). The Bregman and Rudnicky (1975) stimuli, which are shown in Figure 4·3, show how tones can be captured into a stream by having tones that are close in frequency. When A and B were presented by themselves, listeners could easily judge the temporal order. When A and B were flanked by tones F, listeners had a more difficult time. However, if the captor tones C surrounded the flankers, then F streamed with C, A-B split into a different stream, and the listeners could again hear the order of A-B. Thus, if A and B are in the middle of a stream, their order is more difficult to determine.

4.2.2 Continuity illusion

As mentioned above, proximity combined with closure has led to the auditory continuity illusion. In the continuity illusion, sound A seems to continue through sound B, even though sound A is not present during sound B. This illusion works for a tone, or a glide, continuing through noise (Figure 4·2).

A more complex example is shown in Figure 4·4. The top two figures show the two different stimuli that Steiger (1980) presented to listeners. In (b), the broadband noise replaced the glide portion. However, for both the stimuli in (a) and (b), listeners heard the two streams shown in (c) and (d). In (b), a third stream was also heard corresponding to the broadband noise bursts. Thus, the glide complex had been completed, or continued, through the noise. This experiment is important in that the principle of "good continuation" have been overcome by frequency proximity.

Another effect of continuity derives from Bregman and Dannenbring (1973), which is shown in Figure 4·5. In this, listeners were presented with a cyclic pattern of high (H) and low (L) tones, which are either connected (a), or point towards each other (b), or have no trajectory between them (c). Listeners heard one stream in (a) and two streams in (c), but there was a higher probability of hearing one stream over two streams in (b), where they are pointing towards each other. While this can be seen as lending evidence for trajectory mechanisms, it fits in with frequency proximity.

4.2.3 Harmonicity and pitch

Every periodic source has frequency components, called harmonics, at integer multiples of the fundamental frequency, F_0. The subjective experience of F_0 is denoted as pitch, and is influenced by the harmonic content and other attributes of the signal. Consider a speaker producing a vowel at a particular fundamental frequency, e.g.

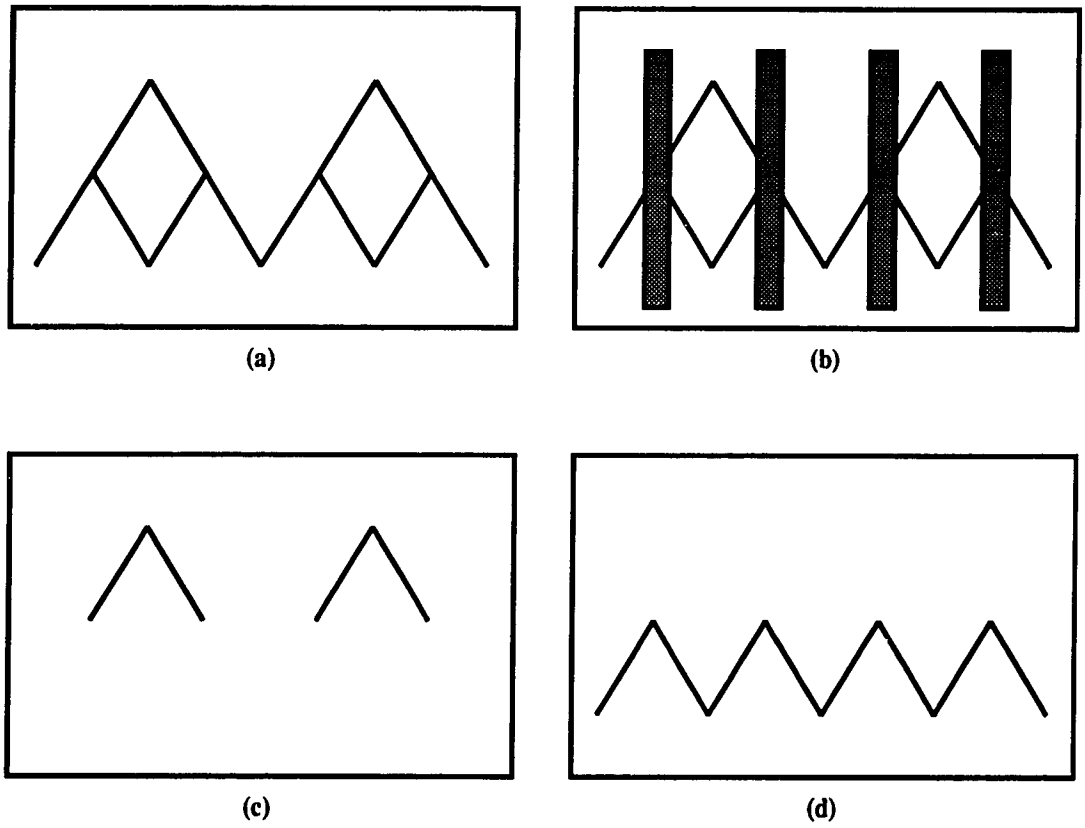


Figure 4·4: Stimuli and percept of the experiment by Steiger (1980). (a) and (b) show the stimuli that were presented to the subjects. In (b), the noise spectra is not added to the glides, but actually replaces the glide portions. For both the stimuli in (a) and (b), listeners hear the two streams shown in (c) and (d). In (b), a third stream is heard corresponding to the broadband noise bursts. After Steiger (1980).

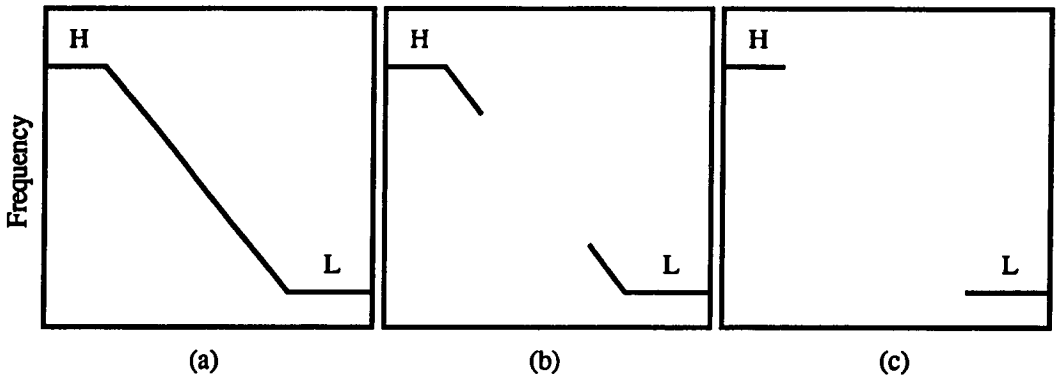


Figure 4-5: Listeners are presented with a cyclic pattern of high (H) and low (L) tones, which are either connected (a), or point towards each other (b), or have no trajectory between them (c). The effect is that listeners heard one stream in (a) and two streams in (c), but that there was a higher probability of hearing one stream in (b), where they are pointing towards each other. After Bregman and Dannenbring (1973).

150 Hz. The vowel contains harmonics at integer multiples, e.g. 300, 450, 600, etc, and the relative amplitudes of these harmonics lead to a given vowel percept. Since a set of related harmonics will correspond to the same source, the pitch can be used to group these harmonic components.

Harmonics of a complex tone can be segregated out from the tone if it is mistuned by 1.5 to 3%, as well as causing the complex pitch to shift. If the mistuning is greater than 3%, the harmonic has little effect on the pitch, and is still heard as a second source (Moore, Glasberg, & Peters, 1985). Also, lower harmonics are easier to capture from a complex than higher harmonics, and harmonics are easier to capture out of a complex if the neighboring harmonics are removed (van Noorden, 1975). Partials spaced 14 semitones apart fuse better than ones that 16 semitones apart (Bregman, 1990). A semitone is the smallest pitch interval in Western music, and two tones separated by a semitone corresponds to tones at frequencies f and (1.06)f. These effects are related to the resolution of the harmonics within the auditory channels

(Cohen, Grossberg, & Wyse, 1994).

Segregation based on harmonicity is used by listeners in speech perception. It has been shown that listeners can use F_0 to segregate multiple voices. Listeners' identification of two concurrent vowels increases as the difference in the two F_0 increases, and plateaus between .5-2 semitones (Scheffers, 1983). When F_0 was an octave apart, the identification was also very poor (Brokx & Noteboom, 1982; Chalika & Bregman, 1989). Since an octave corresponds to a doubling of frequency, half the harmonics for the two vowels will overlap. It should be noted that listeners could identify concurrent vowels with the same F_0 with greater than chance accuracy, implying that listeners can use schema-based segregation. In addition, a formant frequency (frequencies with greater energy that correspond to vowel identity) of a single vowel become segregated when the formant has a differing F_0 (Broadbent & Ladefoged, 1957; Gardner, Gaskill, & Darwin, 1989). Finally, speech stimuli with discontinuous pitch contours tend to segregate at the discontinuities (Darwin & Bethell-Fox, 1977).

While the harmonicity cues can cause components to group, it can compete with frequency proximity cues leading to a bounce or cross percept in the perception of crossing glides.

4.2.4 Bounce and cross percept in crossing glide complexes

The influence of harmonicity is seen in the experiments of Bregman and Doehring (1984), who showed that a glide can be captured into a stream if two partials form a harmonic frame around the glide. While harmonicity can cause streaming, glides which cross produce a bounce percept, presumably due to frequency proximity at the crossing point (Halpern, 1977; Tougas & Bregman, 1990). A bounce percept corresponds to hearing two streams, one with a "∪" shaped percept and another with

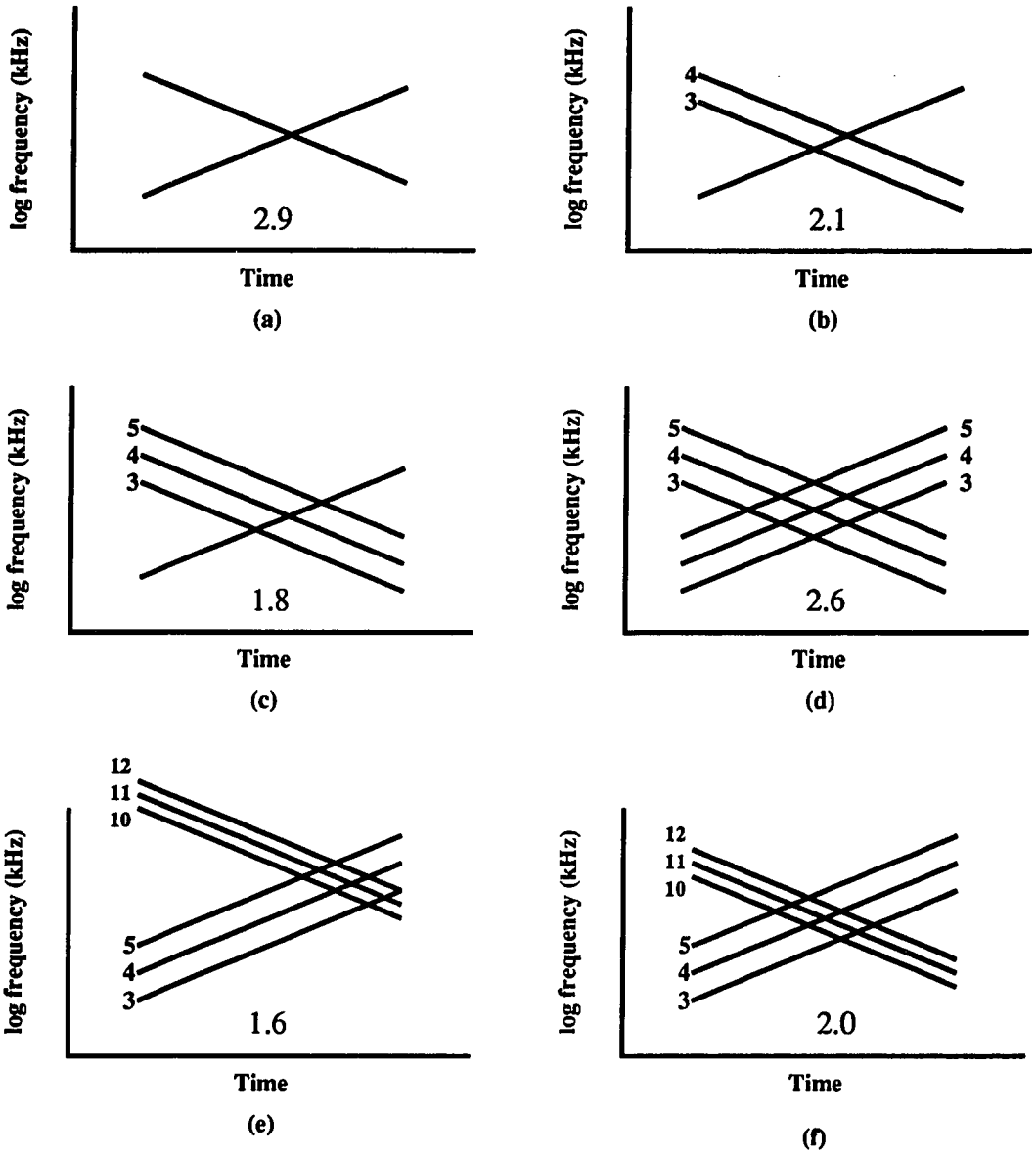


Figure 4-6: Stimuli and listeners' responses in Halpern (1977) for different harmonic conditions. The complex glides were all 1 second long, and the numbers next to a glide is its harmonic number. The numbers below each figure corresponds to the preference of hearing a bounce or a cross: numbers greater than 2.5 correspond to a bounce percept, and numbers below 2.5 correspond to a cross percept. After Halpern (1977).

a "∩" shaped percept, due to the crossing of glides. The cross percept corresponds to hearing two streams, each stream containing one of the glides. Halpern (1977) presented the six different one second glide stimuli shown in Figure 4·6 to subjects and asked them to rate how well they produced a bounce percept. The numbers below each figure corresponds to the preference of hearing a bounce or a cross: numbers greater than 2.5 correspond to a bounce percept, and numbers below 2.5 correspond to a cross percept. The numbers next to the glides correspond to the harmonic number of an underlying F_0. The stimuli in (a) and (d) produced a bounce percept, while the others produced a cross percept. This experiment shows that the harmonic structure in (b) and (c) help to overcome the ambiguity at the crossing point that occurs in (a) and promotes a cross percept.

Tougas and Bregman (1990) performed an experiment very similar to that of Halpern. Tougas and Bregman had four different harmonic stimuli: rich crossing, rich bouncing, all pure, and all rich (Figure 4·7). All but the rich crossing condition produced a bounce percept, even when the interval I was filled with silence, noise, or just the glides. The bounce percept was greatest for rich bouncing, then all pure, and then all rich, for all three interval conditions. The consequence of this experiment is that regardless of noise, silence, or glide during the crossing point, one gets the same percept.

4.2.5 Spatial location

While spatial location seems to be a strong principle for grouping, the auditory system does not treat it as a dominant cue. The principle that frequency components arising from the same spatial location should belong to the same object seems reasonable, but the pliable nature of sound confounds the unambiguous implementation of this idea. Since sounds can travel around objects or corners, one object's sound can

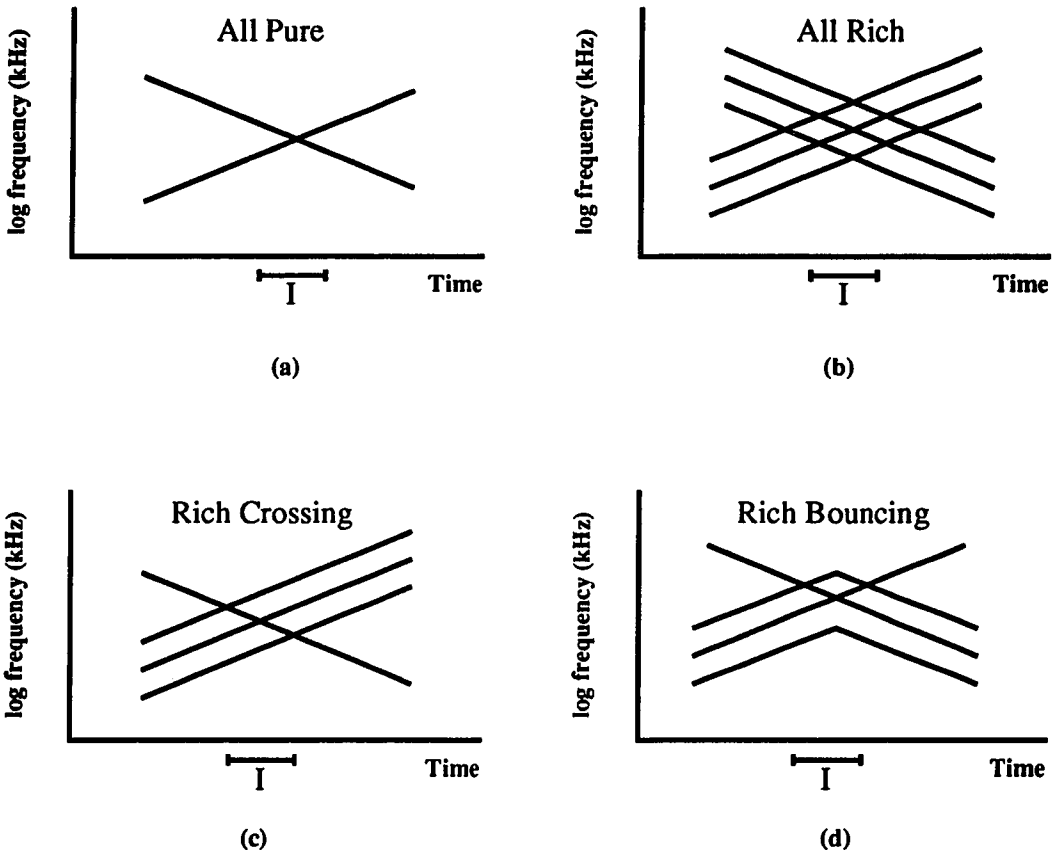


Figure 4·7: Stimuli of Tougas and Bregman (1990) for four different harmonic conditions. All but the rich crossing condition produced a bounce percept, even when the interval I was filled with silence, noise, or just the glides. The order, from greatest to the least, of bounciness was rich bouncing, all pure, and all rich. After Tougas and Bregman (1990).

travel through another object's sound. Moreover, two sounds can arise from the same location, e.g. two talkers on a monophonic radio, which listeners can easily segregate. Thus spatial cues alone are not sufficient to separate streams. Shackleton, Meddis, and Hewitt (1994) presented two different concurrent vowels to listeners and varied the spatial and pitch separation of the two vowels. They found no improvement in identification of both vowels by introducing a spatial difference, while keeping the pitch the same for both vowels. However, by introducing a pitch difference and no spatial cue, performance improved by 35.8%. With both a pitch difference and a spatial difference, the performance improved by 45.5%.

In a free-field environment, there can be up to a 10 dB improvement in intelligibility if the sources are spatially separated (Bronkhorst & Plomp, 1988; Gelfand, Ross, & Miller, 1988). This effect could, however, be due to head shadowing improving the signal-to-noise ratio at one of the ears, and not due to binaural localization per se. However, studies using one spatial lateralization cue, interaural time differences (ITD), over headphones have shown only a slight improvement (4 dB) in intelligibility (Bronkhorst & Plomp, 1988; Carhart, Tillman, & Greetis, 1969; Levitt & Rabiner, 1967).

One piece of evidence that spatial cues effect segregation is binaural masking level difference (BMLD). In this phenomenon, a tone, which is masked by white noise, is presented to both ears, and the level of masking is determined. If a 180 degree phase shift is then induced in one of the tones, then the tone becomes more perceptible, and a new masking level is determined. The difference between the two masking levels is the BMLD. Thus, the ability to perceive the tone was improved by making the tone derive from a different spatial location.

Other experiments have shown that the binaural match between frequency components have to be nearly exact if the auditory system is going to group them based

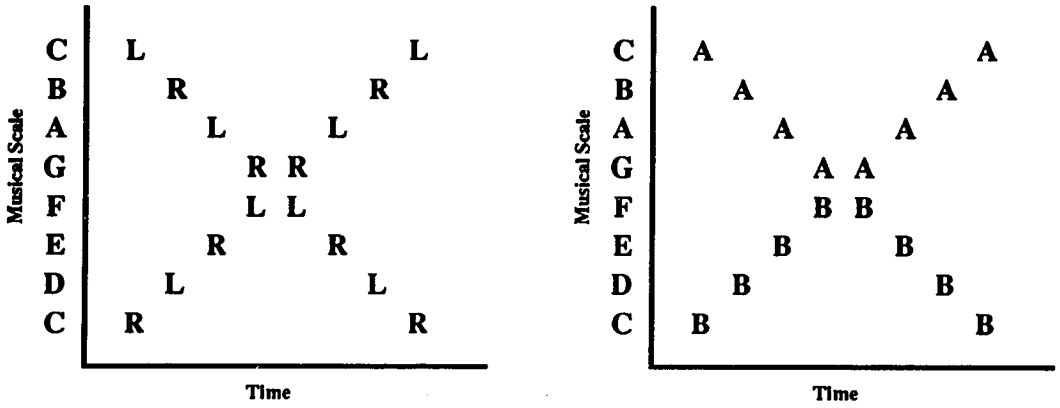


Figure 4-8: (a) Scale illusion in which a downward and an upward scale are being played at the same time, except that every other tone in a given scale is presented to the opposite ear, corresponding to an L or R for left and right ear. (b) The result is that listeners group based on frequency proximity, and heard the two streams A and B. After Deutsch (1975).

on spatial cues. Steiger and Bregman (1982) presented a repeating cycle consisting of a two-tone captor (C) followed by a two-tone target (T) to the left ear. In addition, they presented a 4-tone masker (M) to either the left or right ear, which was synchronous with the target. They manipulated the harmonic structure and the ear to which they presented the masker. They found that binaural fusion of the masker and target occurred when the harmonics of M and T matched exactly. However, if the harmonics of M and T differed (greater than 4% for frequencies less than 1 kHz), the monaural spectral (frequency components) fusion was greater and C streamed with T. Cutting (1976) presented two tokens of /da/ to both ears with an ITD. When the ITD was 4 ms, subjects were more likely to hear two objects. In addition, Cutting showed that if one /da/ was synthesized at 100 Hz and another at 102 Hz, then listeners almost always heard two sources. Thus, the spectral match across the ears has to be quite exact.

Grouping can also affect perceived location. If a tone located in the medial plane

is captured by a left ear tone (due to frequency proximity), as opposed to a right ear tone, then the central tone will be perceived to come from the left side (Bregman & Steiger, 1980). The scale illusion of Deutsch (1975) also illustrates this point (Figure 4·8a). In this illusion, a downward and an upward scale are played at the same time, except that every other tone in a given scale is presented to the opposite ear. In the figure, the ear presentation is shown as an L or R for left and right ear. The result is that listeners grouped the sounds based on frequency proximity, and heard the two streams A and B shown in Figure 4·8b. In addition, right-hand listeners stated that they heard the higher tones (A) in the right ear, and the lower tones (B) in the left ear.

Overall, it seems that spatial cues are secondary cues, and the perceptual system relies more on harmonicity and proximity cues.

4.2.6 Amplitude modulation (AM)

Amplitude modulation (AM) can be a possible cue if the perceptual system groups those frequency components which have correlated amplitude fluctuations. One effect of AM is that the perception of a tone, which is masked by a noise band centered on the tone, can become easier to perceive if another band of noise is modulated with the centered noise (Hall & Grose, 1988). The release of the tone from masking is known as comodulation masking release (CMR). While this effect exists, a recent experiment by Summerfield and Culling (1992) showed that at slow AM rates (2.5Hz), segregation of two vowels did not improved due to AM. So, the influence of AM on segregation of multiple voices of seems unlikely.

4.2.7 Frequency modulation (FM)

Frequency modulation (FM) could act as a streaming cue if the auditory system could detect correlated frequency changes among spectral components. One needs to distinguish coherent FM from incoherent FM. In coherent FM, all partials (a harmonic or inharmonic component of a complex tone) are modulated at the same rate. In incoherent FM, the partials are modulated independently. Changes in F_0 can correspond to coherent FM since all the harmonics are being changed by a proportionate amount. Thus, segregation based on coherent FM could be a result of changes in F_0.

The results from recent psychophysical experiments seem to imply that segregation based on FM is not used. Carlyon (1991) found that with inharmonic complex tone pairs, listeners could not distinguish between coherent and incoherent FM, per se. Extending this, Carlyon (1992) found that if listeners did discriminate between coherent and incoherent FM, it was due to mistuning a harmonic and not to FM explicitly. Moreover, McAdams (1989) showed that by adding vibrato and jitter to different components of three vowel mixture, the components did not segregate. Summerfield (1992) found that identification of a vowel presented with another vowel did not improve when a difference in FM was used, and all the harmonics had been randomly shifted. However, there was some benefit if the components of one vowel in a two vowel presentation was frequency modulated while the other was not (Summerfield & Culling, 1992). This result could be due to pitch difference cues though. Thus, for the most part, it seems that FM is not used as cue for segregation.

4.2.8 Onsets and offsets

Common onset and offset cause grouping, even over sequential grouping (Bregman & Pinker, 1978; Dannenbring & Bregman, 1978). Bregman and Pinker (1978) presented

the stimulus shown in Figure 4·1 as a repeating sequence. They found that as A and B were further separated in frequency, onset and offset synchrony grouped B and C together. However, as B and C became asynchronous, A and B grouped together to form a stream.

The interaction between harmonicity and onset asynchrony was investigated by Darwin and Ciocca (1992) . They found that if a harmonic started 160 ms before rest of a complex tone, then it had a diminished influence on pitch of the complex tone. Moreover, if it started 300 ms before before the complex, then it has no influence on the pitch. Finally, Bregman and Rudnicky (1975) found that two 250 ms tones that have 88% overlap fuse into one stream.

The effect of onset asynchrony has also been investigated in speech perception. If a tone is added near the first formant of a synthetic vowel, causing the first formant to shift, it leads to a different vowel percept. The original vowel percept can be restored if the tone has a 30 ms onset asynchrony, implying that the tone and the original vowel were in separate streams (Darwin, 1984; Darwin & Sutherland, 1984). Darwin and Sutherland presented a 56 ms vowel, synthesized at a fundamental frequency of 125 Hz. They found that if the harmonic at 500 Hz started 240 ms earlier (Figure 4·9a), then it had a diminished contribution to the vowel identity. If a 1000 Hz tone was then added that started at the same time as the 500 Hz harmonic and stopped at the vowel (Figure 4·9b), then the harmonic's contribution increased slightly. Thus, the addition of a harmonic of 500 Hz, namely the 1000 Hz tone, caused them to be grouped together and thereby disinhibiting the contribution of the harmonic during the complex tone. This manipulation showed that the reduced contribution of the harmonic was not due solely to adaptation. Roberts and Moore (1991) extended this observation to show that the tone can be harmonic or inharmonic.

While not as strong as onset asynchrony, offset asynchrony influences grouping. A

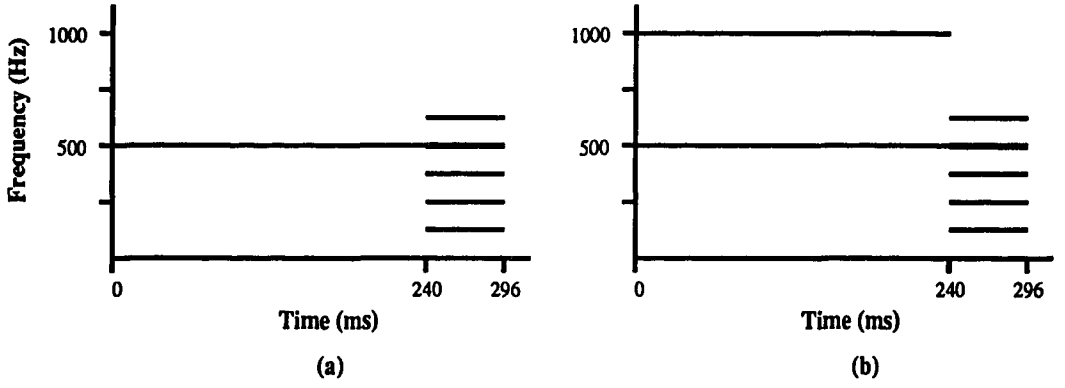


Figure 4·9: (a) If a harmonic at 500 Hz started 240 ms before the rest of a short synthetic vowel, then it has a diminished contribution to the vowel identity. (b) If a 1000 Hz tone was added that started at the same time as the 500 Hz harmonic and stopped at the vowel, the harmonic's contribution increases slightly due to the grouping of the 500 and 1000 Hz tones. After Darwin and Sutherland (1984).

harmonic which has an offset asynchrony of 30 ms with respect to a vowel complex contributes less to its identity than one with a synchronous offset (Darwin, 1984; Darwin & Sutherland, 1984).

4.3 Existing models of segregation

Meddis and Hewitt (1992) presented a static model that segregated concurrent vowels based on pitch. The pitch was derived using an autocorrelation. However, the model did not handle temporally-varying stimuli. Brown (1992) and Cooke (1991) have presented complex models which perform segregation of temporally-varying stimuli. These models use pitch cues, derived from autocorrelation methods, to perform segregation. However, these models use time-frequency kernels to achieve segregation. In other words, they treat the stimuli as a static pattern, a spectrogram, and then perform dynamic programming and spatio-temporal processing, which treats time as another spatial dimension. None of these models has tried to model the process dynamically.

4.4 Model of auditory streaming and grouping

The neural model developed in this thesis suggests how harmonicity and frequency proximity interact in the brain. The model, which is shown in Figure 4·10, consists of several stages. The model first preprocesses the incoming signal in the peripheral processing modules. The preprocessed signal is then used to group frequency components based on pitch.

The first several stages are based on the physiology and psychophysics of the auditory periphery (Cohen, Grossberg, & Wyse, 1992, 1994). The peripheral processing preemphasizes the signal, or boosts the amplitude of higher frequencies, which emulates the outer and middle ears. Next, the preemphasized signal is filtered by a bank of bandpass filters, which emulates the cochlea. Finally, an energy measure is obtained at the output of these filters.

This energy measure feeds into the different fields in the spectral stream layer, where different fields correspond to different streams. There is competition between these sheets for each frequency component. No component can be simultaneously allocated to two streams after the competition acts. In addition, this competition causes a component that is not harmonically related to the other components in a given stream to "pop out" of the spectrum assigned to that stream and become active in another stream. The spectral stream layer has reciprocal connections with the pitch stream layer to determine which spectral components belong to a given pitch. Thus, a pitch is associated with each active stream. The feedback from the pitch stream layer reinforces consistent components and suppresses inconsistent components, as in Adaptive Resonance Theory (Grossberg, 1980; Carpenter & Grossberg, 1991). Therefore, the listener's percept corresponds to the activity at the spectral stream layer when there is resonance between it and the pitch stream layer.

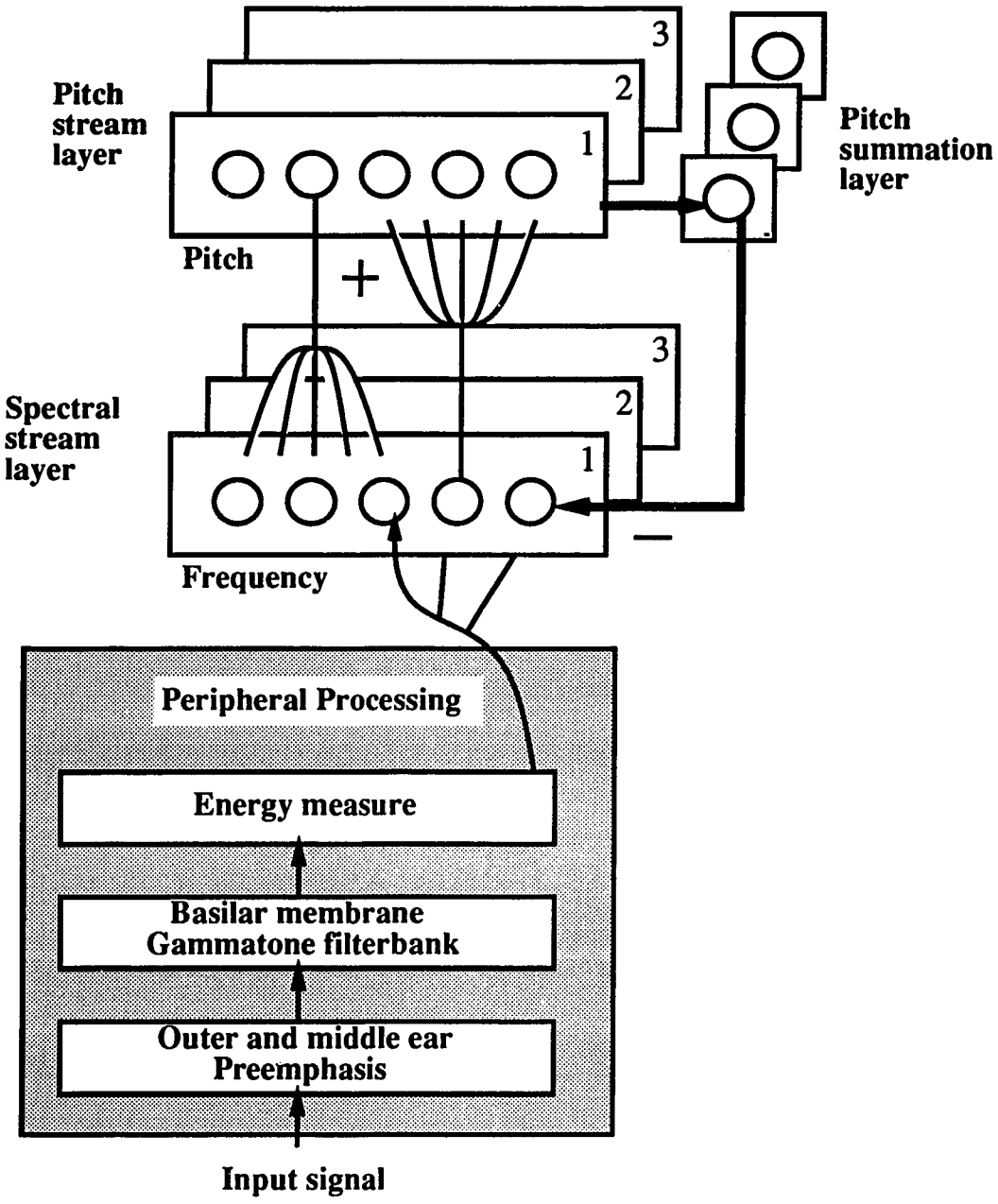


Figure 4·10: Block diagram of the auditory streaming model.

4.4.1 Auditory peripheral processing

Outer and middle ear

The outer and middle ear act as a broad bandpass filter, linearly boosting frequencies between 100 to 5000 Hz. An approximation to this is to preemphasize the signal using a simple difference equation:

$$y(t) = x(t) - A * x(t - \Delta t), \tag{4.1}$$

where A is the preemphasis parameter, and Δt is the sampling interval. In the simulations, A was set to 0.95, and $\Delta t = 0.125$ ms, corresponding to a sampling frequency of 8 kHz.

Cochlear filterbank

The overall effect of the basilar membrane is to act as a filterbank, where the response at a particular location on the basilar membrane acts like a bandpass filter. This bandpass characteristic has been modeled as a fourth order gammatone (de Boer & de Jongh, 1978; Cohen et al., 1994) filter:

$$g_{f_0}(t) = \begin{cases} t^{n-1} e^{-2\pi t * b(f_0)} \cos(2\pi f_0 t + \phi) & t > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{4.2}$$

and its frequency response is:

$$G_{f_0}(f) = [1 + j(f - f_0)/b(f_0)]^n, \tag{4.3}$$

where n is the order of the filter, f_0 is the center frequency of the filter, ϕ is a phase factor, and $b(f)$ is the gammatone filter's bandwidth parameter, corresponding to:

$$b(f) = 1.02ERB(f). \tag{4.4}$$

The equivalent rectangular bandwidth (ERB) of a gammatone filter is the equivalent bandwidth that a rectangular filter would have if it passed the same power:

$$ERB(f) = 6.23e^{-6}f^2 + 93.39e^{-3}f + 28.52. \tag{4.5}$$

Sixty gammatone filters, which were equally spaced in ERB, were used to cover the range 100 Hz to 2000 Hz. The output of each gammatone filter was converted into an energy measure.

Energy measure

The energy measures a short-time energy spectra (Cohen et al., 1992, 1994):

$$e_f(t) = \frac{\Delta t}{W} \sum_{k=0}^{W/\Delta t} |g_f(t - k\Delta t)|^2 e^{-\alpha \Delta t k}, \tag{4.6}$$

where $e_f(t)$ is the energy measure output of the gammatone filter $g_f(t)$ centered at frequency f at time t, W is the time window over which the energy measure is computed, and α represents the decay of the exponential window. In the simulations, $\alpha = 0.995$, and W = 5 ms. The output of the energy measure feeds identically to the multiple fields in the spectral stream layer.

4.4.2 Spectral stream layer

Segregation based on harmonicity is achieved by having objects compete for frequency channels, which are excited by their pitch counterparts and supported by the bottom-up input (Figure 4·11). The spectral stream layer is a plane with one axis representing frequency, and the other axis representing different auditory streams.

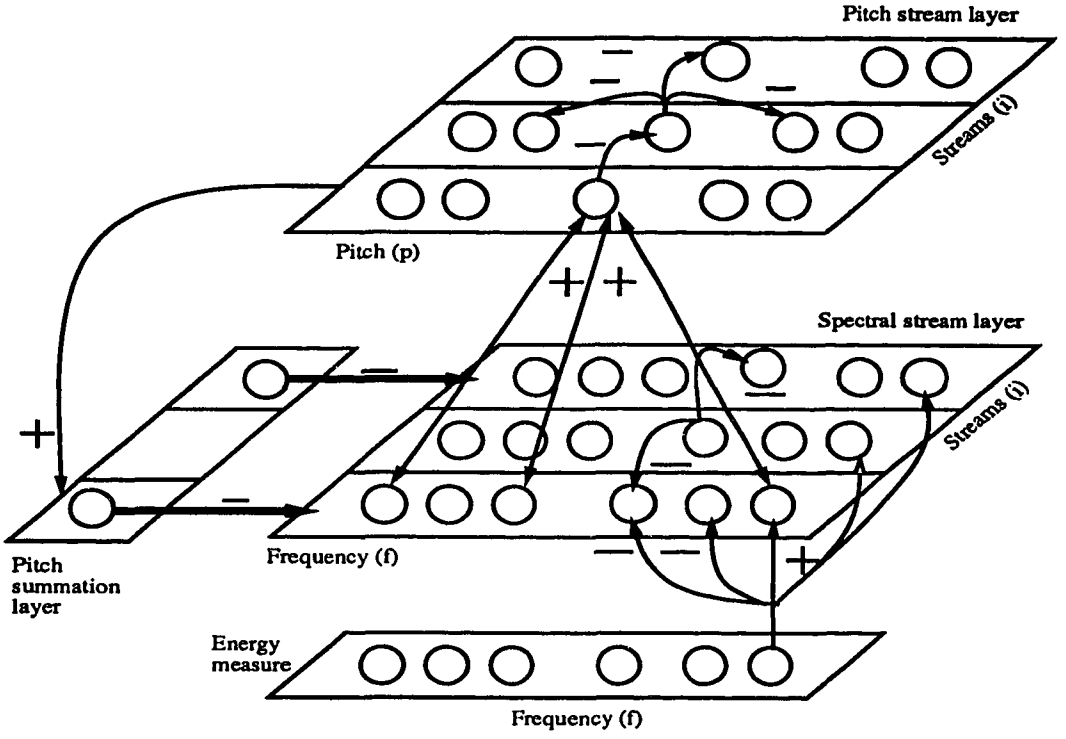


Figure 4·11: Interaction between the energy measure, the spectral stream layer, the pitch stream layer, and the pitch summation layer. The energy measure layer is fed forward in a frequency-specific one-to-many manner to each frequency-specific stream node in the spectral stream layer. In addition, this feed-forward activation is contrast-enhanced. There is also competition within the spectral stream layer across streams for each frequency so that a component is allocated to only one stream at a time. Each stream in the spectral stream layer activates its corresponding pitch stream in the pitch stream layer. Each pitch neuron receives excitation from its harmonics in the corresponding stream. Since each pitch stream is a winner-take-all network, only one pitch can be active at any given time. Across streams in the pitch stream layer, there is asymmetric competition for each pitch so that one stream is biased to win and the same pitch can not be represented in another stream. Finally, the winning pitch neuron feeds back excitation to its harmonics in the corresponding spectral stream. The stream also receives non-specific inhibition from the pitch summation layer, which sums up the activity at the pitch stream layer for that stream. This non-specific inhibition helps to suppress those components that are not supported by the top-down excitation, which plays the role of a priming stimulus or expectation (Carpenter & Grossberg, 1991).

Each frequency channel in the energy measure, e_f, feeds up to each stream's corresponding frequency channel in the spectral stream layer S_f in a one-to-many manner, so that all streams in the spectral stream layer receive equal bottom-up excitation. After the spectral stream layer becomes activated, the different streams activate their corresponding pitch streams in the pitch stream layer. When a pitch is selected in a given stream, it feeds back excitation to its spectral harmonics, and inhibits that pitch value in other streams in the pitch stream layer. In addition, non-specific inhibition, via the pitch summation layer, helps to suppress components that do not belong to the given pitch within its stream.

The following equation describes the dynamics of the spectral stream layer:

$$\dot{S}_{if} = -AS_{if} + [B - S_{if}]\mathcal{E}_{if} - [C + S_{if}]\mathcal{I}_{if} \tag{4.7}$$

$$\mathcal{E}_{if} = \sum_g D_{fg}s(e_g) + F\sum_p \sum_k M_{f,kp}g(P_{ip})h(k) \tag{4.8}$$

$$\mathcal{I}_{if} = \sum_{g\neq f} E_{fg}s(e_g) + J\sum_{k\neq i} \sum_g N_{fg}[S_{kg}]^+ + LT_i \tag{4.9}$$

where S_{if} is the activity of the spectral stream layer neuron corresponding to the ith stream and frequency f. Term $-AS_{if}$ in (4.7) is the spontaneous decay. Term $D_{fg}s(e_g)$ in (4.8) is the excitation from the energy measure, which has been passed through a sigmoid $s(x)$ to compress the dynamic range:

$$s(x) = \begin{cases} x^2/(N_s + x^2), & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.10}$$

Similarly, $E_{fg}s(e_g)$ in (4.9) is the inhibition from the energy measure, which has been passed through a sigmoid $s(x)$. Thus, with both $D_{fg}s(e_g)$ and $E_{fg}s(e_g)$, each spectral stream layer receives a contrast-enhanced version of the energy measure.Both D_{fg} and E_{fg} are Gaussians which are centered at frequency f, and have standard

deviation parameters, σ_D and σ_E, and scaling parameters D and E, respectively:

$$D_{fg} = DG(f, \sigma_D) = D\frac{1}{\sigma_D\sqrt{2\pi}}e^{-.5(f-g)^2/\sigma_D^2} \tag{4.11}$$

$$E_{fg} = EG(f, \sigma_F) = E\frac{1}{\sigma_E\sqrt{2\pi}}e^{-.5(f-g)^2/\sigma_E^2} \tag{4.12}$$

In addition, the term $F\sum_p\sum_k M_{f,kp}g(P_{ip})h(k)$ in (4.8) is the sum of all the pitches p which have a harmonic kp near frequency f in the pitch stream layer corresponding to stream i. In 4.8, $g(x)$ is a sigmoid function:

$$g(x) = \begin{cases} x^2/(N_g + x^2), & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.13}$$

h(k) is the harmonic weighting function, which weights the lower harmonics more heavily than higher harmonics:

$$h(k) = \begin{cases} 1 - M_h log_2(k), & \text{if } 0 < M_h log_2(k) < 1 \\ 0, & \text{else} \end{cases} \tag{4.14}$$

and $M_{f,kp}$ is a normalized Gaussian, so that if a harmonic is slightly mistuned it will still be within the Gaussian and thus get partially reinforced. The width of the Gaussian dictates the tolerance of mistuning. Kernel $M_{f,kp}$ is centered at frequency f and has a standard deviation parameter, σ_M:

$$M_{f,kp} = G(f, \sigma_M) = \frac{1}{\sigma_M\sqrt{2\pi}}e^{-.5(f-kp)^2/\sigma_M^2} \tag{4.15}$$

The term $J\sum_{k\neq i}\sum_g N_{fg}[S_{kg}]^+$ in (4.9) represents the competition across streams for a component, so that a harmonic will belong to only one object. This inhibition embodies the principle of "exclusive allocation." Since a harmonic can be mistuned slightly, a Gaussian window N_{fg} exists within which the competition takes place.

Kernel N_{fg} is centered at frequency f and has a standard deviation parameter, σ_N:

$$N_{fg} = G(f, \sigma_N) = \frac{1}{\sigma_N\sqrt{2\pi}} e^{-.5(f-g)^2/\sigma_N^2} \tag{4.16}$$

Term LT_i in (4.9) is the inhibition from the pitch summation layer, which non-specifically inhibits all components in stream i. The effect of this is to subtract out those non-harmonic components which are not reinforced by the top-down excitation from the pitch unit in the pitch stream layer. This is akin to the matching process used in Adaptive Resonance Theory (Carpenter & Grossberg, 1991; Grossberg, 1980). As a result of this matching process, a spectral stream layer neuron can become:

- Active if only an energy input is present (bottom-up automatic activation),
- Inactive if only a pitch input is present (top-down priming),
- Active if both energy and pitch inputs are present (bottom-up and top-down consistency),
- Inactive if both energy and pitch inputs are present, but the spectral component is not a harmonic of pitch (bottom-up and top-down inconsistency).

The first constraint allows bottom-up activation to initiate the segregation process. So, if there is no pitch unit that is active, then there is no inhibition from the pitch stream layer, via the pitch summation layer. Thus, the spectral stream layer will become active. The second constraint makes sure that the pitch units do not activate spurious spectral units by themselves, but only in conjunction with an input. This is accomplished by letting the inhibition from the pitch summation layer be no smaller than the excitation from the pitch units. The third and fourth constraints state that only harmonics of the particular pitch which are present in the input are excited. This is accomplished by setting the combined excitation from the input and pitch

stream unit to be greater than the inhibition from the pitch summation layer. If a spectral unit is a harmonic of a pitch P and it has an input at that frequency, then the spectral unit will remain active. However, if the unit is not a harmonic (or a slightly mistuned harmonic), then the inhibition from the pitch summation layer will be greater than only the bottom-up input. In all the simulations, the parameters were set to: $A = 1, B = 1, C = 1, D = 500, E = 450, F = 3, J = 1000, L = 5, M_h = .3, N = .01, N_s = 10000, N_g = .01, \sigma_D = .2, \sigma_E = 4, \sigma_M = .2$, and $\sigma_N = 1$.

4.4.3 Pitch summation layer

The pitch summation layer sums up the pitch activity at stream i, and provides inhibition LT_i to stream i's spectral stream layer in (4.7)-(4.9) so that only those harmonic components that correspond to the selected pitch remain active:

$$\dot{T}_i = -AT_i + [B - T_i] \sum_p g(P_{ip}) \tag{4.17}$$

where $g(x)$ is the sigmoid function described above. In the simulations, $A = 100, B = 100$.

4.4.4 Pitch stream layer

To determine the pitch, the neural network pitch model of Cohen, Grossberg, and Wyse (1992, 1994), called the SPINET model, was used. The original pitch model had two components: the spectral layer and a pitch layer. The spectral and pitch representations have been modified so that there are multiple streams such that competition occurs between pitch units within and across streams. The modified pitch strength activation is:

$$\dot{P}_{ip} = -AP_{ip} + [B - P_{ip}]\mathcal{E}_{ip} - [C + P_{ip}]\mathcal{I}_{ip} \tag{4.18}$$

$$\mathcal{E}_{ip} = E \sum_k \sum_f M_{f,kp}[S_{if} - \Gamma]^+ h(k) \tag{4.19}$$

$$\mathcal{I}_{ip} = J \sum_{p \neq q} H_{pq} g(P_{iq}) + L \sum_{k>i} g(P_{kp}), \tag{4.20}$$

where P_{ip} is the pth pitch unit of object i. The term $E \sum_k \sum_f M_{f,kp}[S_{if} - \Gamma]^+ h(k)$ in (4.19) corresponds to the Gaussian excitation $M_{f,kp}$ from the spectral layer which have suprathreshold components near a harmonic kp of pitch p, which is weighted by the harmonic weighting function $h(k)$. The harmonic weighting function $h(k)$ and the Gaussian $M_{f,kp}$ are same as in the spectral layer (eq. 4.14 and 4.15, respectively). The term $J \sum_{p \neq q} H_{pq} g(P_{iq})$ in (4.20) represents the symmetric off-surround inhibition across pitches within a stream. The off-surround competition across pitches within a stream makes the layer act as a winner-take-all so that only one pitch tends to be active within a stream. In addition, H_{pq} is defined to be one within a neighborhood around pitch unit j and zero otherwise, so that a stream can maintain a pitch even if the pitch fluctuates.

$$H_{pq} = \begin{cases} 1, & \text{if } \mid p - q \mid > \sigma_H \\ 0, & \text{else} \end{cases} \tag{4.21}$$

The term $L \sum_{k>i} g(P_{kp})$ in (4.20) represents asymmetric inhibition across streams for a given pitch, so that only one stream will activate a given pitch. This asymmetry across streams also provides a systematic choice of streams, and prevents deadlock between two streams for a given pitch, since all pitch streams receive equal bottom-up excitation from the spectral layer initially. In all the simulations, the parameters were set to: $A = 100, B = 1, C = 10, E = 5000, J = 300, L = 2, \sigma_H = .2$, and $\Gamma = .005$.

4.5 Simulation results of model

The model is here shown to qualitatively emulate bounce percepts for crossing glides, as well as the continuity illusion. Figure 4·12 shows the stimuli and the listeners' percepts that the model emulates. It should be reiterated that the percept that a listener would hear corresponds to the resonant activity in the spectral layer.

4.5.1 Inharmonic simple tones

If two inharmonic tones are presented, then they should segregate into two different streams since they do not have a common pitch (Moore et al., 1985). Figure 4·12a shows the stimulus and the listeners' percept for two inharmonic tones. Figure 4·13a shows the spectrogram for two inharmonic tones, whose frequencies are 358 Hz and 1233 Hz. Figure 4·13b shows the result after peripheral processing, i.e. the result after the energy measure. Figure 4·14 shows the resulting spectral and pitch layers for the two tone stimulus for two different streams. The figures show that initially the streams compete for the tones, but the first stream, which is inherently biased in the pitch stream layer, wins the higher frequency component, allowing the second stream to capture the lower frequency tone. This figure also shows that the higher frequency tone

Figure 4·15 shows a schematic of how the grouping process works for the two inharmonic tones. After the two tones are processed by the peripheral processing, the higher frequency tone has a larger activity due to the preemphasis. The preprocessed activities feed into the spectral stream layers at time t = 0. Since there is no top-down activity at the spectral stream layers, the two spectral layers are equally active. Next, at time t = t1, the pitch stream layer receives activation from the spectral stream layer. Since stream 1's pitch layer is inherently biased over stream 2's pitch layer, and since the higher frequency tone has a larger activity, the 1233 Hz tone is chosen

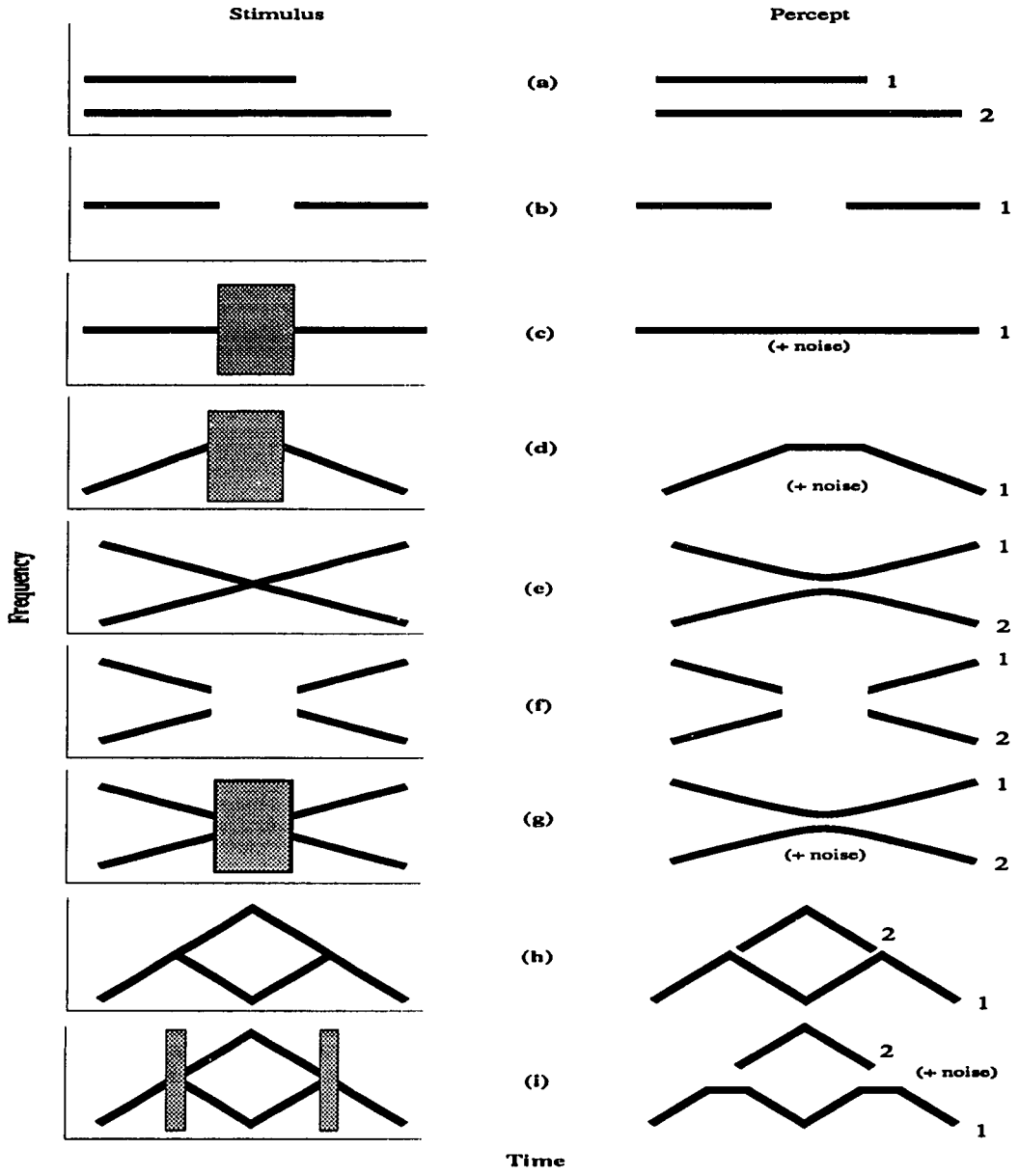


Figure 4·12: Stimuli and the listeners' percepts that the model is capable of emulating. The hashed boxes represent broadband noise. The stimuli consist of: (a) two inharmonic tones, (b) tone-silence-tone, (c) tone-noise-tone, (d) a ramp or glide-noise-glide, (e) crossing glides, (f) crossing glides where the intersection point has been replaced by silence; (g) crossing glides where the intersection point has been replaced by noise, (h) Steiger (1980) diamond stimulus, and (i) Steiger (1980) diamond stimulus where bifurcation points have been replaced by noise.

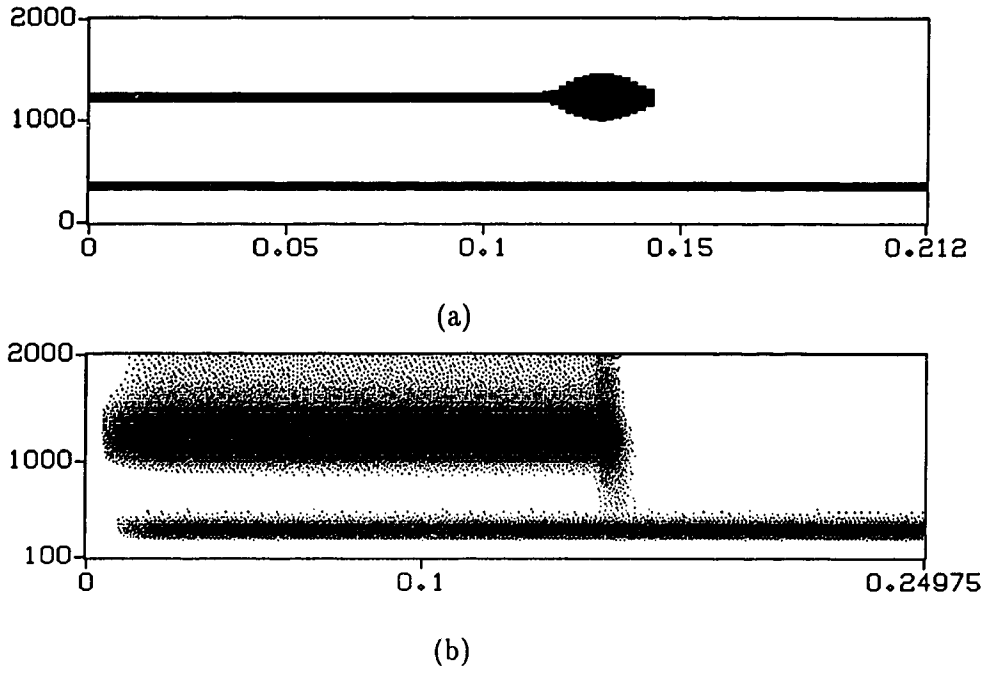
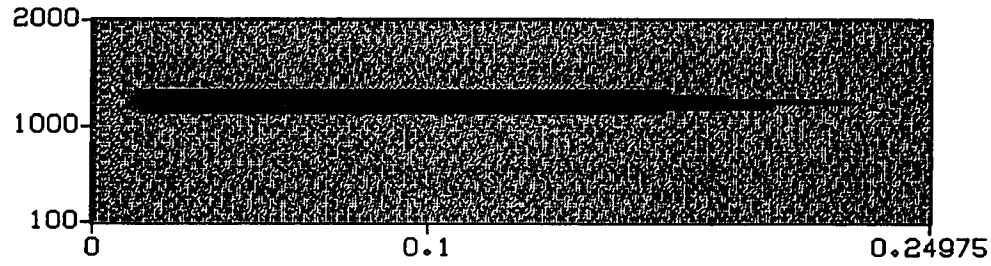
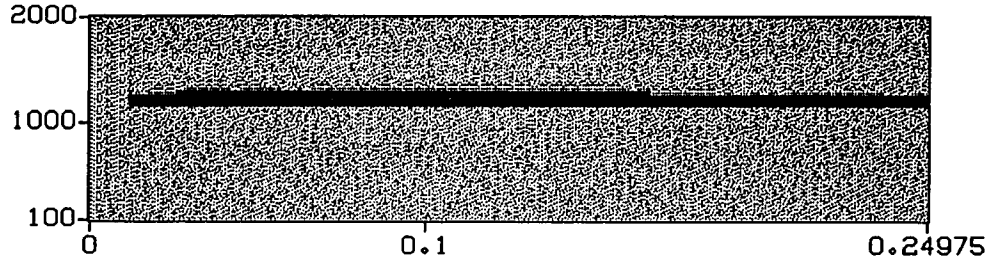


Figure 4·13: (a) spectrogram and (b) result of energy measure for the two tone stimulus.

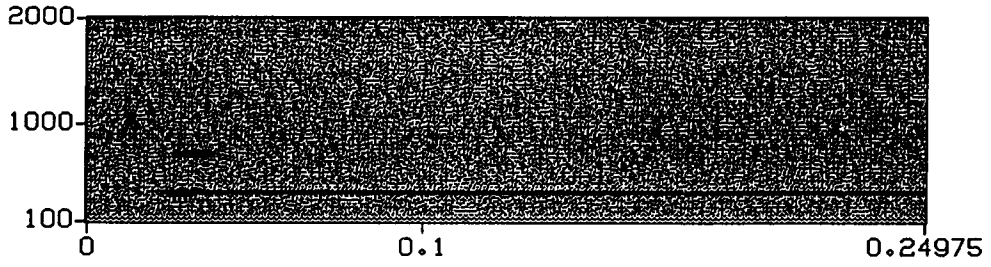
by stream 1's pitch layer. In addition, since the pitch layer is a winner-take-all network, only one pitch can be active within a pitch stream layer. Once the 1233 Hz tone is chosen by stream 1, the corresponding frequency in stream 2's pitch layer is inhibited by the stream 1's winning pitch neuron, allowing the 358 Hz tone to be captured by stream 2's pitch layer. Next, at time t = t2, the winning pitch neurons excite their corresponding harmonic components in the spectral layer. In addition, the non-specific inhibition (shown as the darker arrow) inhibits all components in the spectral layer. Therefore, those components which are not specifically excited by the pitch layer will be suppressed. For example, the 358 Hz tone is suppressed in stream 1 since it is receiving top-down non-specific inhibition and no top-down specific excitation; whereas the 1233 Hz tone receives top-down excitation allowing it to remain active.



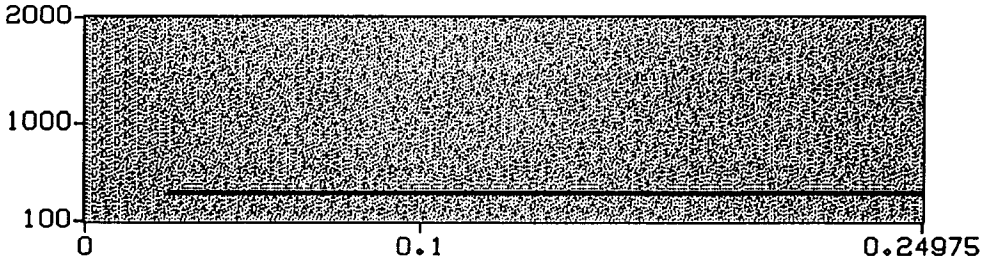
(a)



(b)



(c)



(d)

Figure 4·14: Model results for the two tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

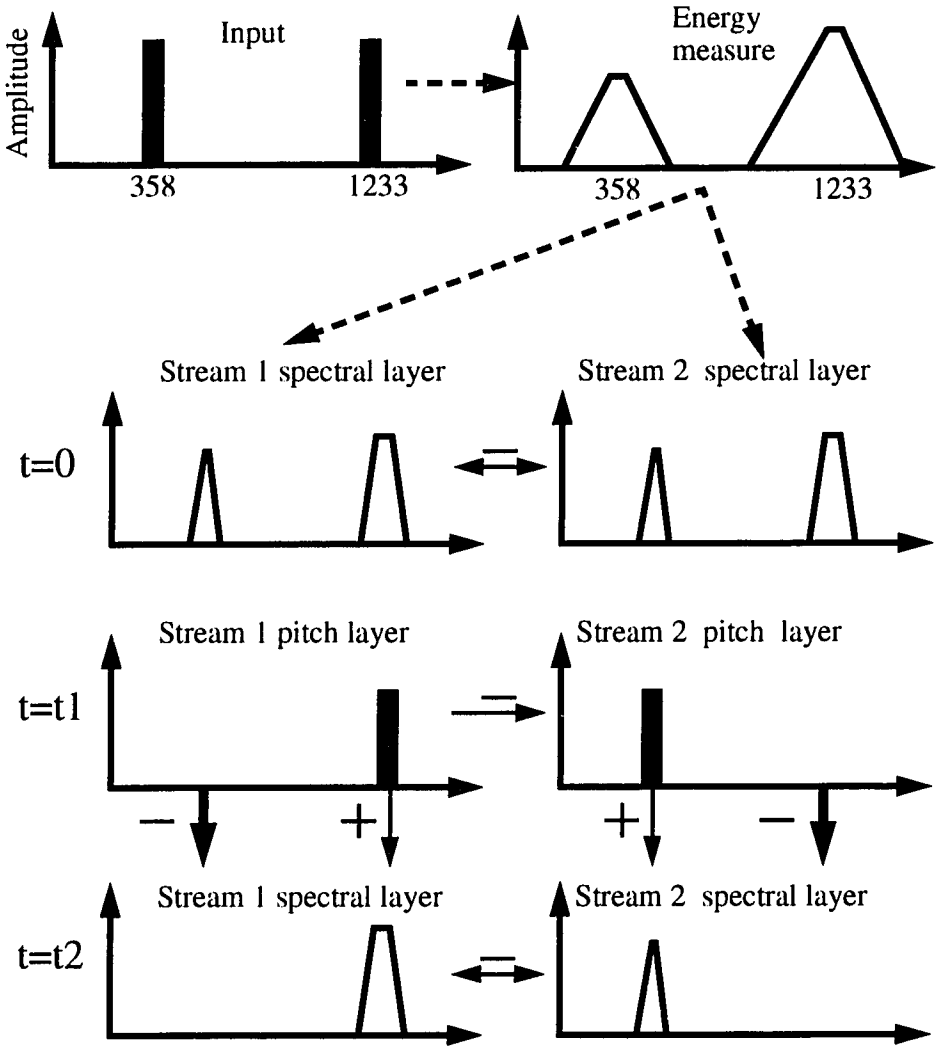


Figure 4-15: Schematic of how the model segregates the two inharmonic tones into two different streams. See text for explanation.

4.5.2 Continuity illusion

The model is capable of producing the continuity illusion: continuation of a tone in noise, even though the tone is not physically present in the noise (Miller & Licklider, 1950). In order to appreciate the result for tone-noise-tone condition, one should consider the result of the model for a tone-silence-tone stimulus (Figure 4·12b). For this stimulus, the tone should not continue across the silence, but should stop at the onset of silence. Figure 4·16 shows the spectrogram and the result after the peripheral processing for the tone-silence-tone stimulus. Figure 4·17 shows the resulting spectral and pitch layers for the tone-silence-tone stimulus for two different streams. The figures show that the first stream captures the tone, which decays into to the silent interval but does not remain active in the silent interval. Since the model does not have any onset/offset mechanisms, the spectral stream activity slowly decays into the silent interval. The same stream then captures the tone after the silence as well. The second stream is not active since there are no extraneous components to capture.

Now, consider the case where the silent interval has been replaced by noise, i.e. the tone-noise-tone stimulus. For this stimulus, the tone percept should continue across the noise, even though the tone is not physically present during the noise interval. Figure 4·18 shows the spectrogram and the result after the peripheral processing for the tone-noise-tone stimulus. Figures 4·19 shows the resulting spectral and pitch layers for the stimulus for the first two streams, and Figure 4·20 shows a third stream. The figures show that the first stream captures the tone, which then continues through and past the noise interval.

The reason that the tone continues through the noise derives from two factors. The first factor is that the spectral layer slowly integrates the input, and so, the noise is temporally averaged, or smoothed over time. Due to this smoothing, if there is no top-down activity, the noise is relatively constant over time. The second factor is that

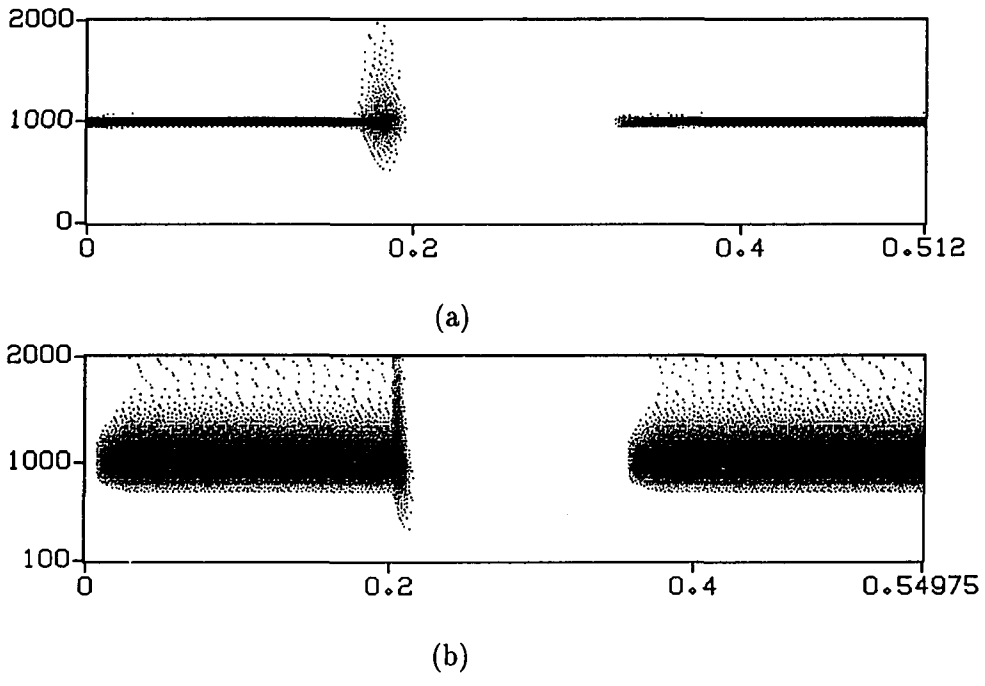
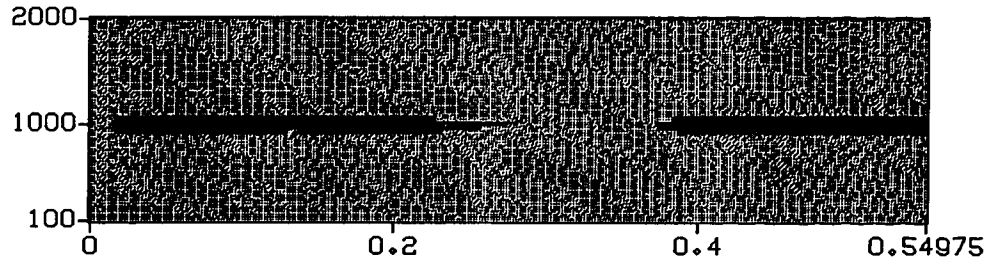
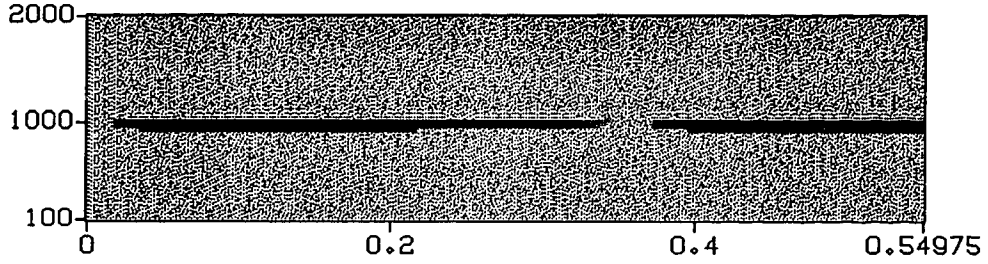


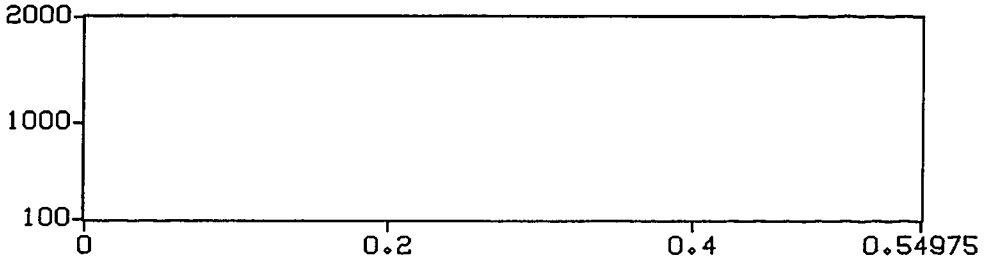
Figure 4·16: (a) spectrogram and (b) result of energy measure for the tone-silence-tone stimulus.



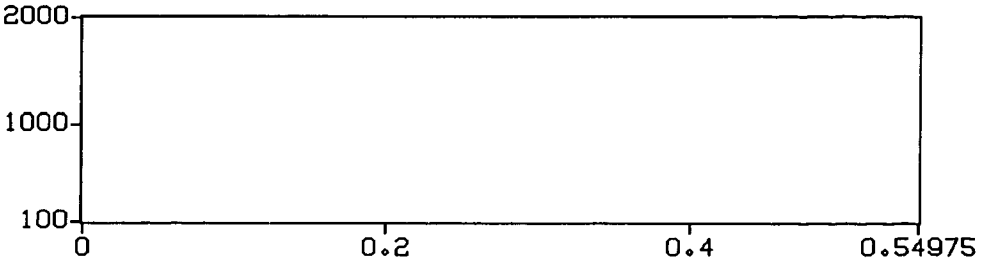
(a)



(b)



(c)



(d)

Figure 4·17: Model results for the tone-silence-tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

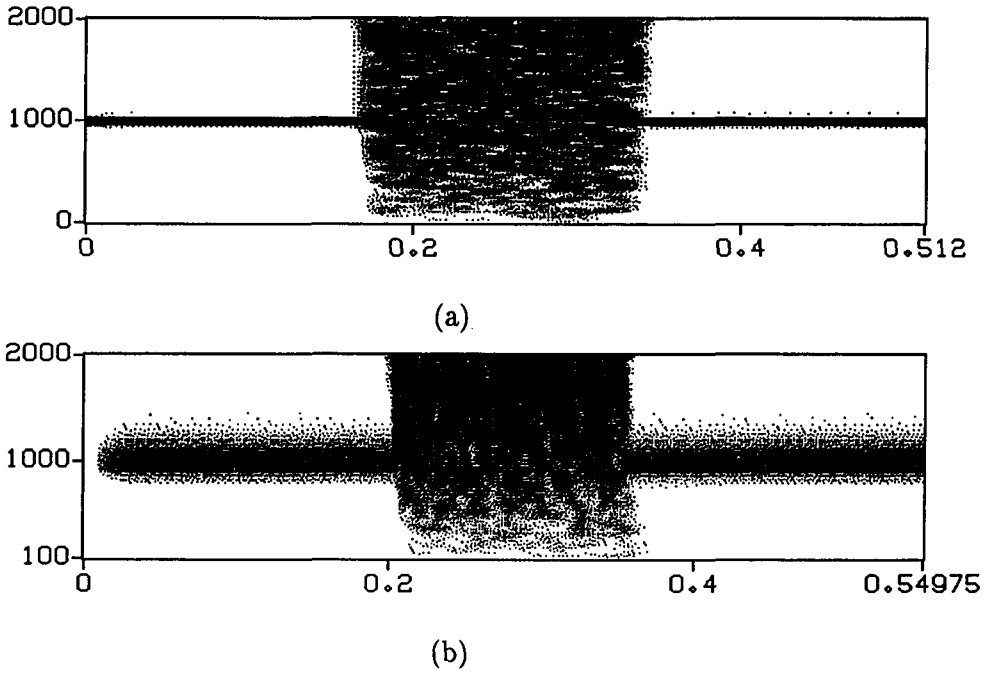


Figure 4·18: (a) spectrogram and (b) result of energy measure for the tone-noise-tone stimulus.

the top-down activity from the pitch layer remains active at the onset of the noise due to the prior tone. Due to both of these factors, the noise at the same frequency as the tone is reinforced by the top-down activity, while the other frequency components are inhibited, allowing the "tone" to complete across the noise. The second and third streams contain the other spurious noise. The reason that the second stream captures the high frequency noise as opposed to the low frequency noise is due to preemphasis: the noise at the highest frequency is most active, and so it is captured by the second stream. If more streams were present in the model, then they would contain other noise components.

Similar to the tone-noise-tone stimulus, the model is capable of producing the continuity illusion for the ramped stimulus shown in Figure 4·12d. Figure 4·21 shows the spectrogram and the result after the peripheral processing. Figures 4·22 shows the

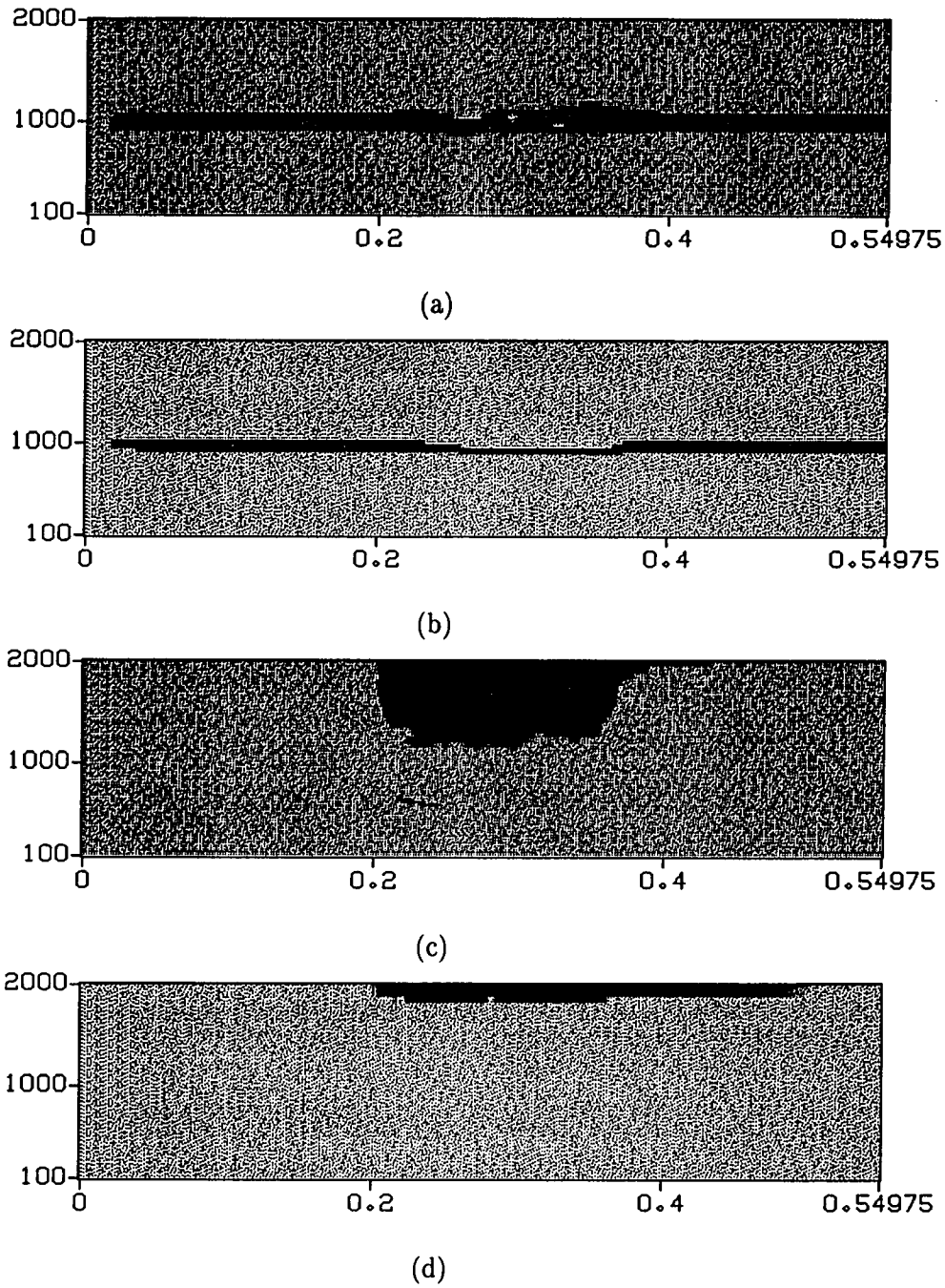


Figure 4-19: Model results for the tone-noise-tone stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

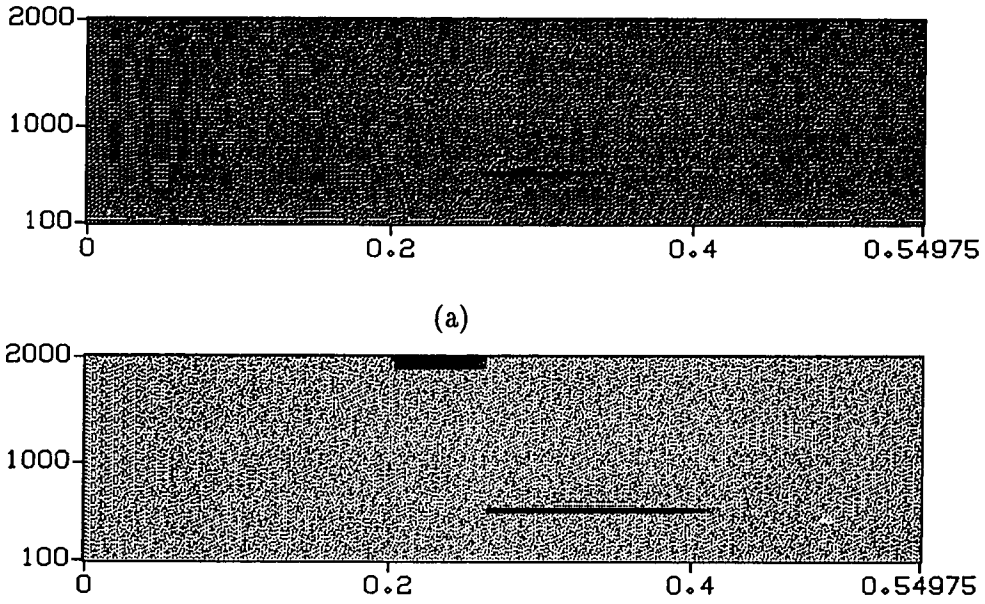


Figure 4·20: The (a) spectral and (b) pitch stream layers for stream 3 for the tone-noise-tone stimulus.

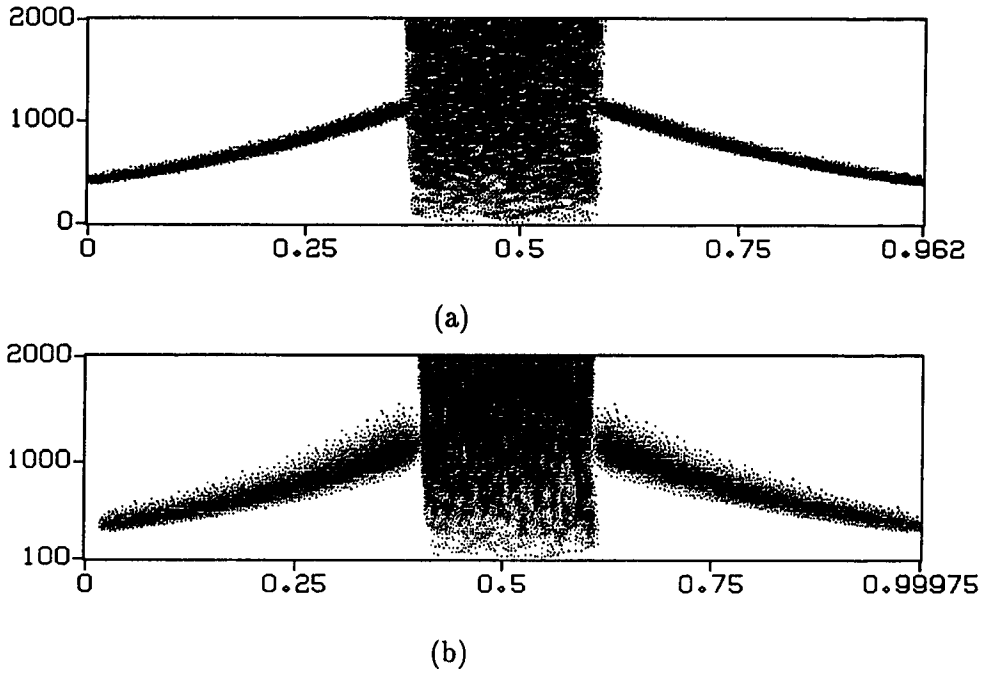


Figure 4·21: (a) spectrogram and (b) result of energy measure for the ramp stimulus.

resulting spectral and pitch layers for the stimulus for the two different streams. The figures show that the first stream captures the upward glide, which then continues through the noise interval. After the noise interval, the same stream captures the downward glide, leading to the ramp percept. The reason that the ramp completes across the noise is due to the same reason that the tone completes across the noise in the tone-noise-tone stimulus; namely, the temporal averaging at the spectral stream layer and the prior top-down excitation from the pitch stream layer. Also, during the noise interval, some noise adjacent to the plateau is active since the top-down inhibition is not strong enough to suppress this activity. Meanwhile, the second stream contains the extraneous noise. If other streams were present, they would also capture some noise components.

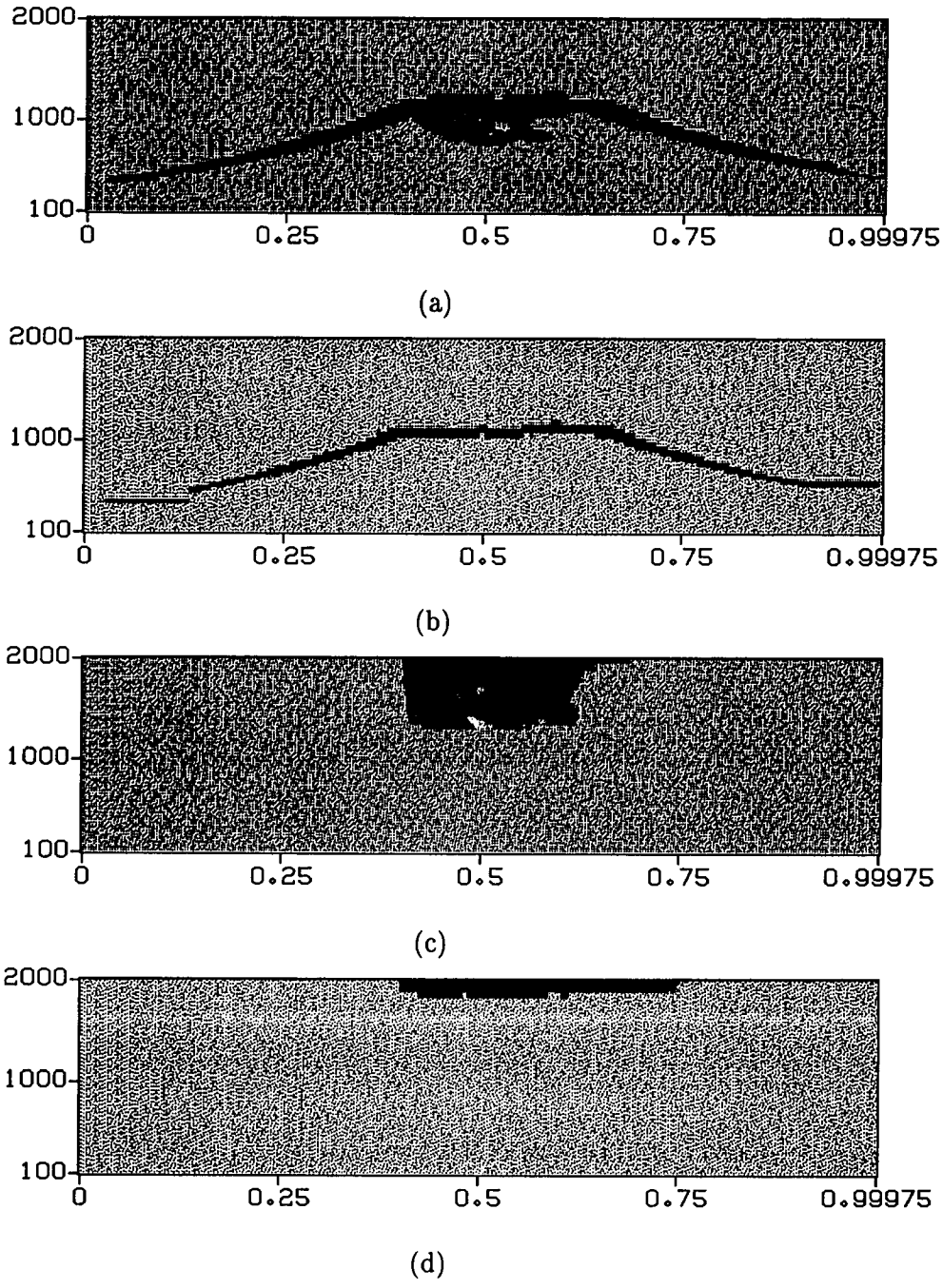


Figure 4·22: Model results for the ramp stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

4.5.3 Bounce percepts for crossing glides

The model is capable of qualitatively replicating the Halpern (1977) and the Tougas and Bregman (1990) data. For these stimuli, one obtains bounce percepts for crossing glides (Figure 4·12e), even if the crossing interval is replaced by silence (Figure 4·12f) or noise (Figure 4·12g). Figure 4·23 shows the spectrogram and the result after the energy measure for the standard crossing glide stimulus; and Figure 4·24 shows the resulting spectral and pitch activity for the two streams. As one can see, one stream contains the "∪" percept, while the other stream has a "∩" percept. The reason one obtains the bounce percept for the standard crossing glide stimulus is due to the following. Initially, the higher frequency glide is captured by the first stream since it has a larger activation, and thus the lower frequency glide is captured by the second stream. The glides are maintained within their streams as they approach the intersection point. At the intersection point, the glides activate multiple, adjacent channels at the spectral layer. These adjacent channels can belong to the two different streams such that the larger frequency channel belongs to the first stream, and thus, grouped with the upper glide; and the lower adjacent frequency channel belongs to the second stream, and thus, grouped with the lower glide.

Figure 4·25 shows the crossing glide stimulus for the silent-center condition and the result of the energy measure. Figure 4·26 shows the spectral and pitch layers for two different streams. The result corresponds to a bounce percept, which does not continue across the silent interval. The reason one obtains the grouping of the upper glides derives from the following. The first stream captures the higher frequency glide at the onset of the stimulus and after the silent interval since these component have a larger activity than the lower frequency glides due to preemphasis. Since these components have a larger activity, the first stream will choose these components, leading to the grouping of the upper glides by stream 1, and the lower glides by

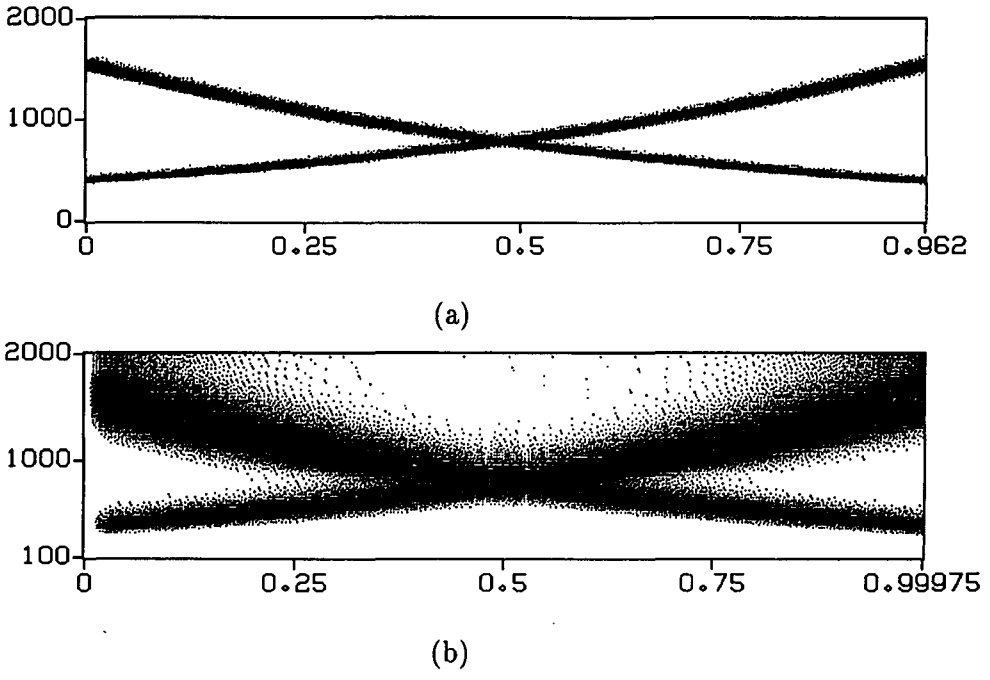


Figure 4·23: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus.

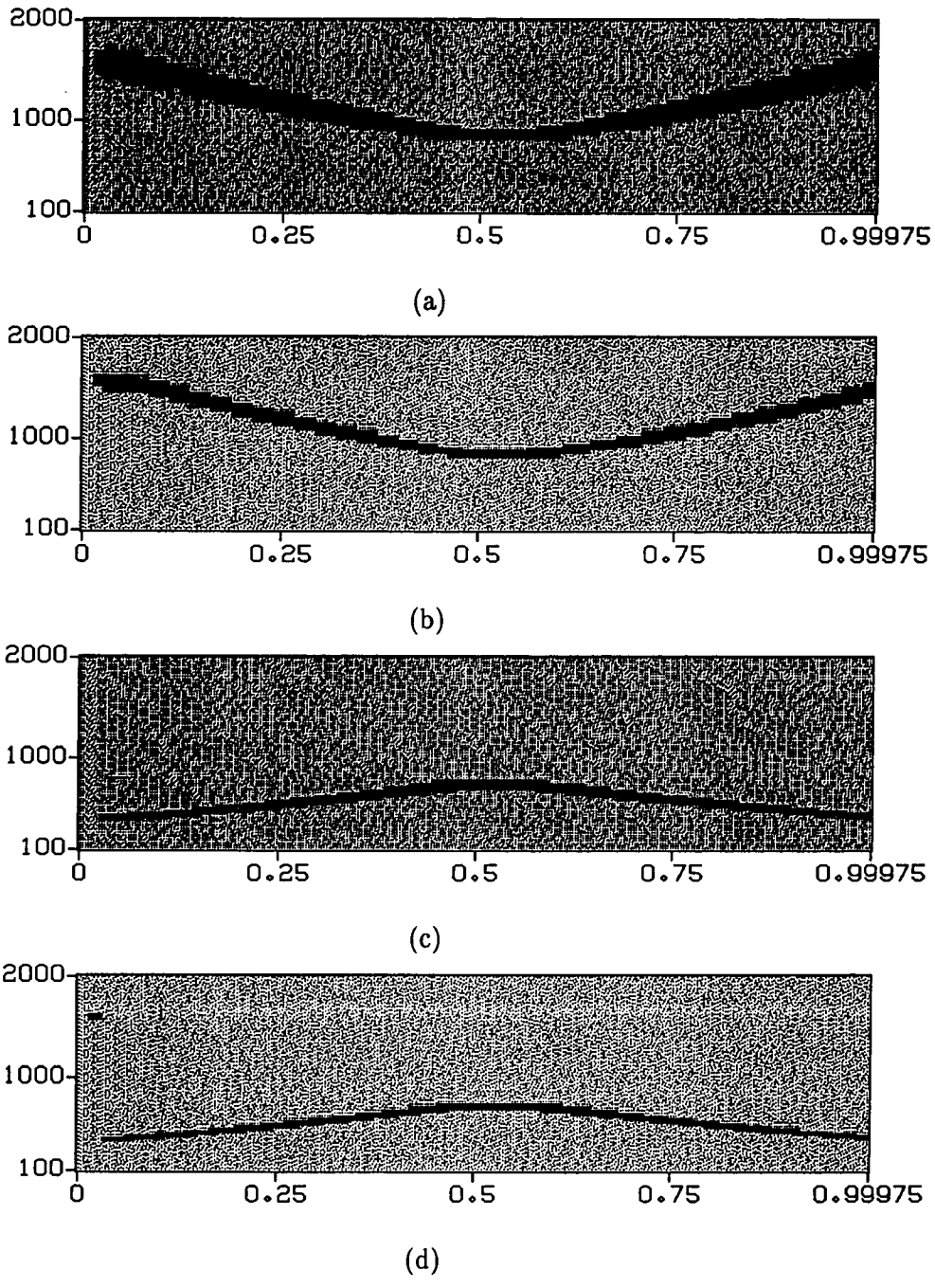


Figure 4·24: Model results for the crossing glide stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

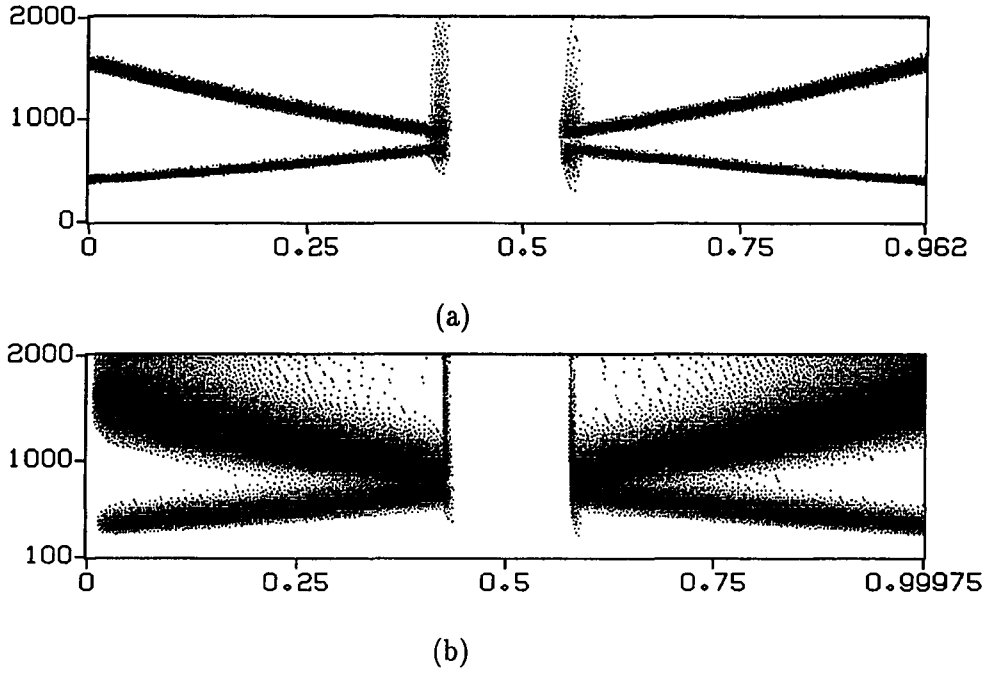


Figure 4·25: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with silence replacing the intersection point.

stream 2; i.e. a bounce percept.

Figure 4·12g shows the crossing glide stimulus where the intersection point has been replaced by noise, and the subjects' percepts of a bounce that is completed across the noise interval. Figure 4·27 shows the spectrogram and the result of the energy measure for the crossing glide with noise-center stimulus, and Figure 4·28 shows the spectral and pitch layers for two different streams. Once again, the bounce percept is evident, but there is continuity of the bounce through the noise interval. Stream 2 shows some noise activity that "leaks" through, which is due to not enough top-down inhibition. The reason that the model produces the bounce phenomenon derives from the results of the continuity illusion and the standard crossing glide stimulus. Initially, the upper frequency glide is chosen by stream 1, and the lower frequency glide is chosen by stream 2, just as in the standard crossing glide stimulus.

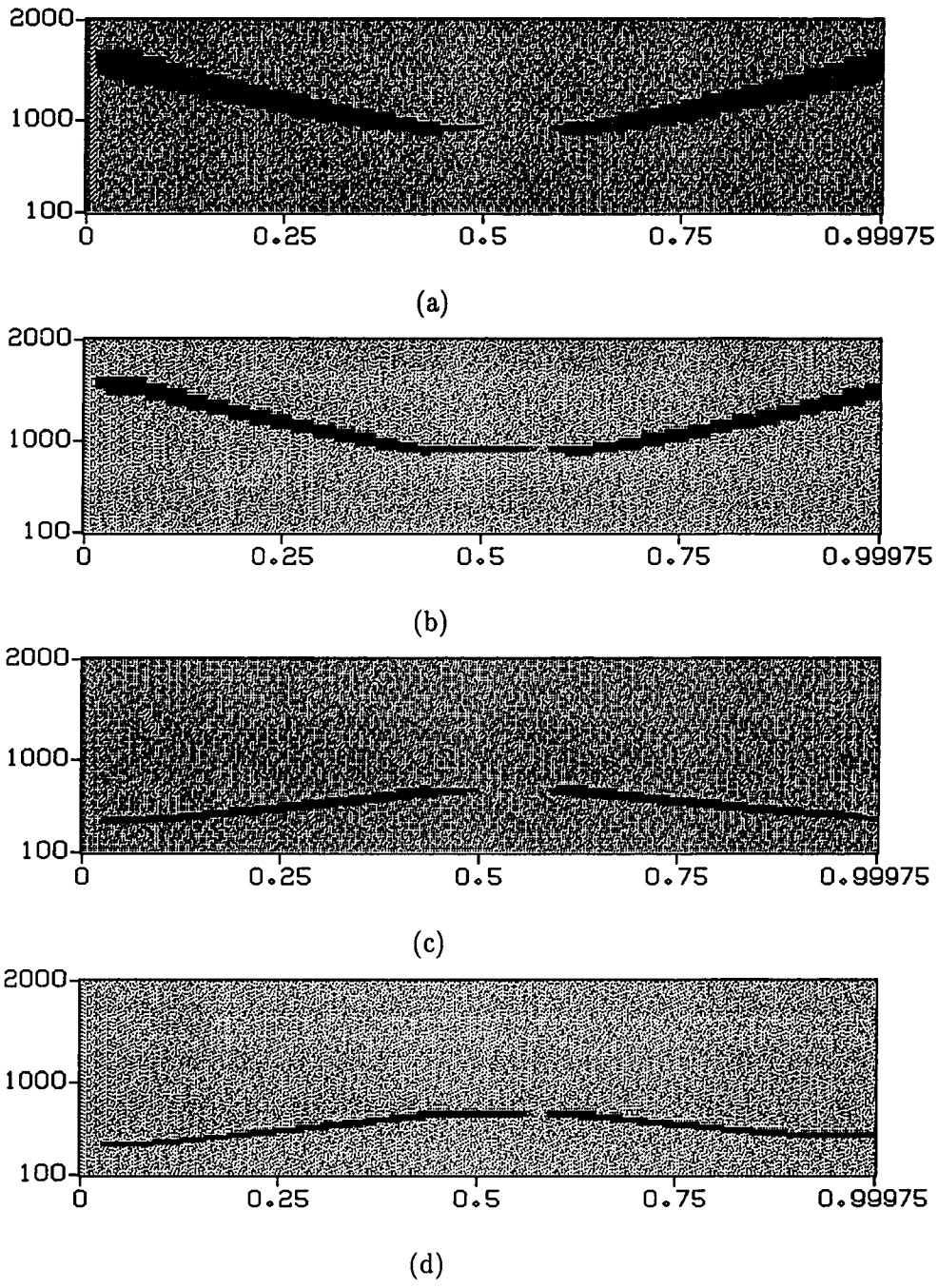


Figure 4·26: Model results for the crossing glide stimulus with silence replacing the intersection point. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

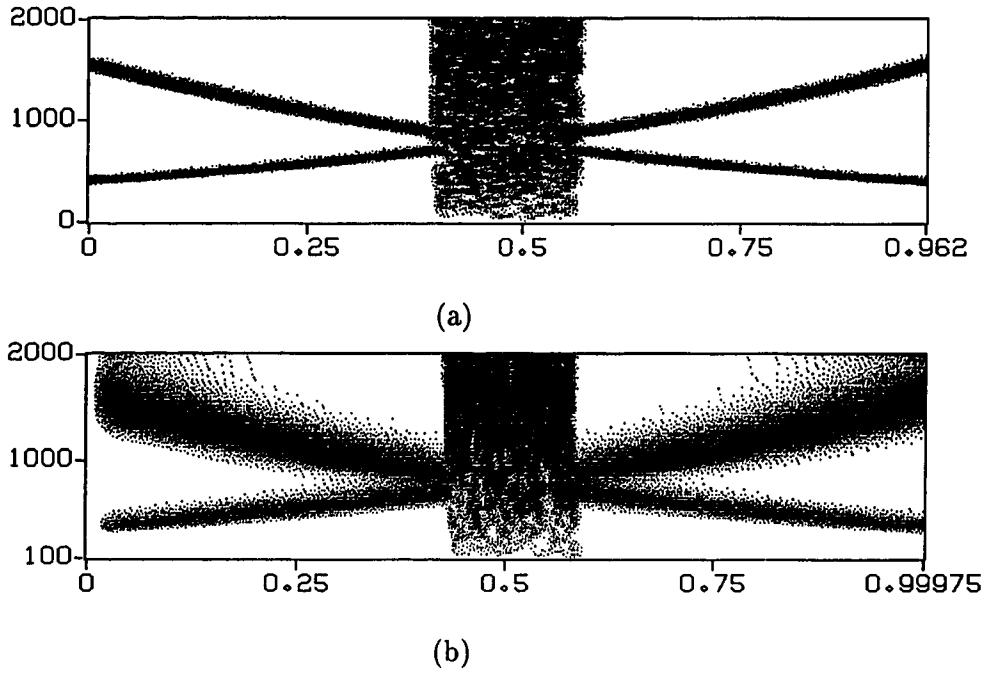


Figure 4·27: (a) spectrogram and (b) result of energy measure for the crossing glide stimulus with noise replacing the intersection point.

The continuity illusion explanation, e.g. the ramp stimulus, applies during the noise interval. At the onset of the noise, the top-down activity from the pitch layer helps maintain the "tone" across the noise interval at the same frequency as the offset of the glide. In addition, the temporal averaging of the noise at the spectral stream layer provides uniform activity over time that aids the resonance between the spectral and pitch layers, and thus, maintaining the "tone" across the noise interval. At the offset of the noise, the glides are at approximately the same frequency as the "tones" that were continuing through the noise. Thus, these glides are grouped with the stream that has a "tone" close to its frequency. As a result, one obtains a bounce percept, where the bounce completes across the noise interval.

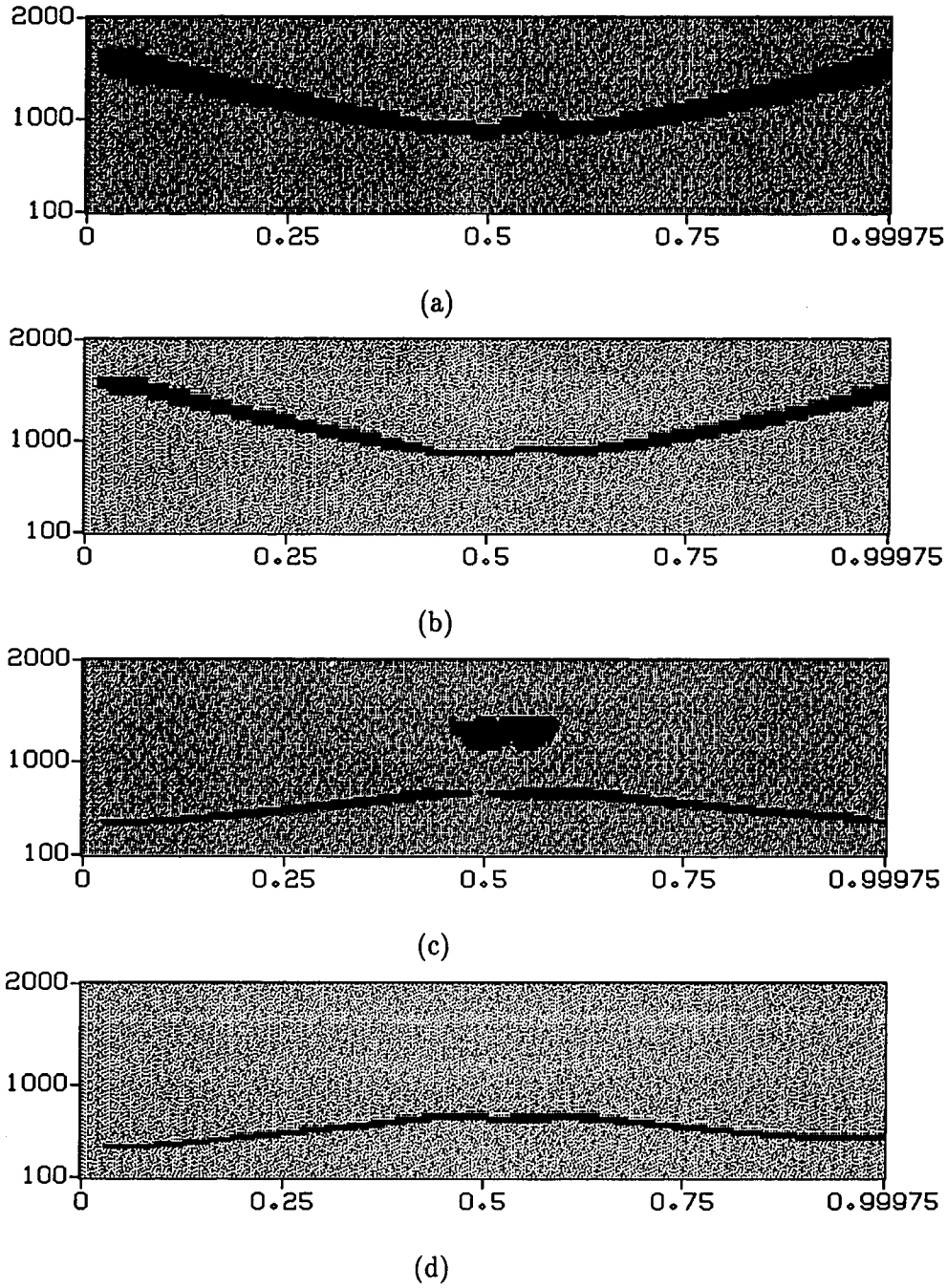


Figure 4-28: Model results for the crossing glide stimulus with noise replacing the intersection point. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

4.5.4 Steiger (1980) diamond stimulus

For the Steiger (1980) diamond stimulus (Figure 4·12h), the percept consists of two streams, a "M" stream and an inverted "V" stream. This percept shows that the principle of continuity can be overcome by frequency proximity. Figure 4·29 shows the Steiger (1980) stimulus and the result after the peripheral processing. Figure 4·30 shows the spectral and pitch layer for two different streams. As one can see, the lower "M" shaped component falls into one stream, while the inverted "V" is in the other stream, which qualitatively emulates the percept. The reason the model emulates the Steiger data is similar to the explanation for the bounce percept for the standard crossing glide explanation. Initially, stream 1 is active with the lower frequency glide and stream 2 is inactive, since there is only one component present in the stimulus. At the bifurcation point, stream 1 continues with the lower frequency glide since this frequency component was previously active in stream 1. In other words, due to the temporal averaging of the spectral layer activity and resonance with the pitch layer, the frequency component that was activated immediately prior to the bifurcation point will remain active and group with the same frequency component immediately after the bifurcation point. Since the first stream groups the lower frequency glides together, the second stream is capable of capturing the higher frequency glides. Thus, stream 1 contains the "M" percept, while stream 2 contains the inverted "V" percept.

Figure 4·31 shows the spectrogram and the result of the energy measure for the Steiger (1980) stimulus where the bifurcation points have been replaced by noise. Figure 4·32 shows the spectral and pitch layers for the two streams for the Steiger (1980) stimulus when the bifurcation points have been replaced by noise. The figures show that the "M" and the inverted "V" segregate into two different streams, and the "M" continues across the noise interval. The noise activates other streams, which are not shown. The reason the model emulates this percept derives from the explanation

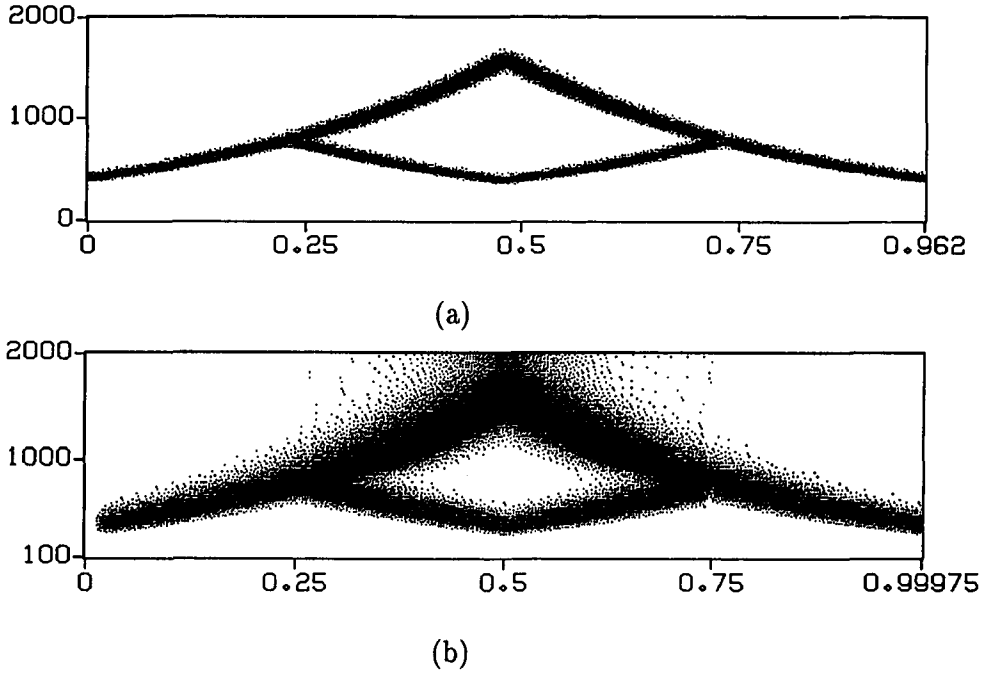
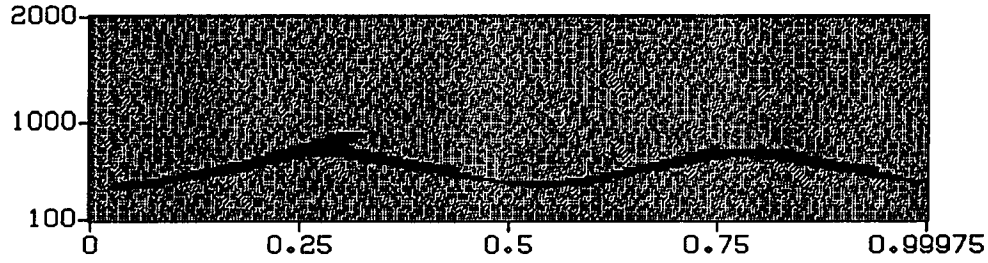
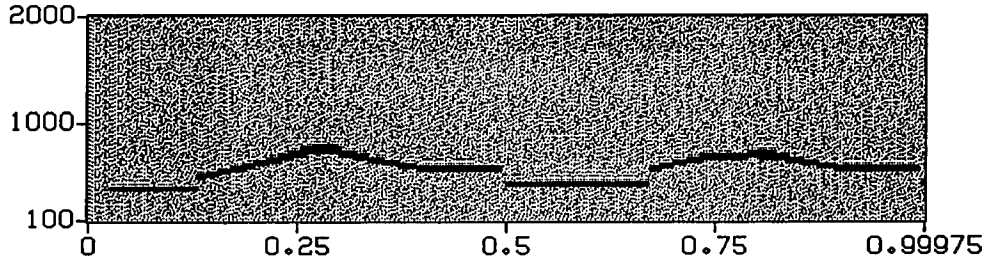


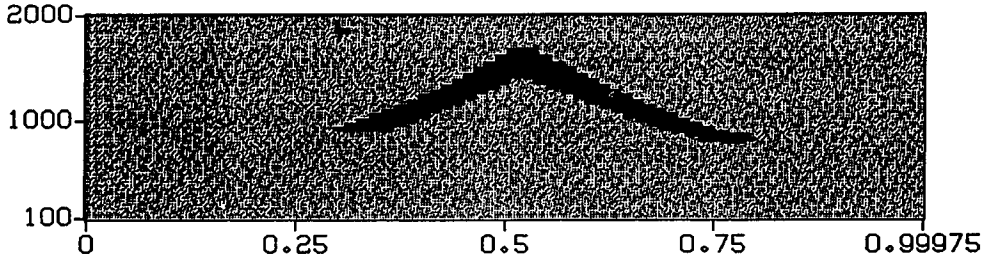
Figure 4·29: (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus.



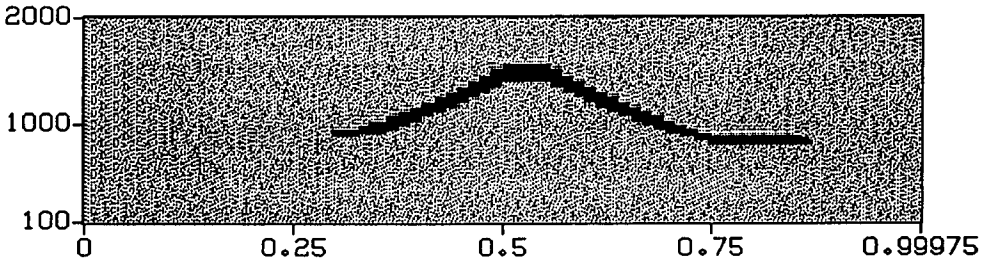
(a)



(b)



(c)



(d)

Figure 4-30: Model results for the Steiger (1980) diamond stimulus. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

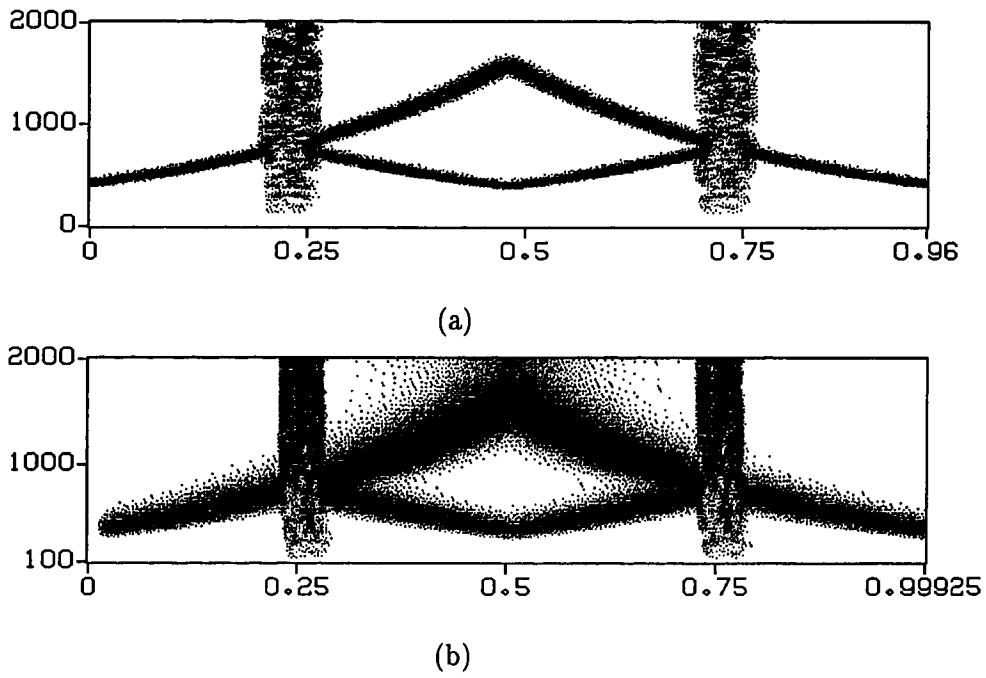


Figure 4-31: (a) spectrogram and (b) result of energy measure for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points.

of the Steiger (1980) diamond stimulus and the continuity illusion, e.g. the ramp stimulus. Stream 1 initially captures the increasing glide, while stream 2 is inactive, just as in the Steiger (1980) diamond stimulus. During the noise interval, stream 1 completes across the noise interval just as in the ramp stimulus, allowing stream 2 to capture the inverted "V" component.

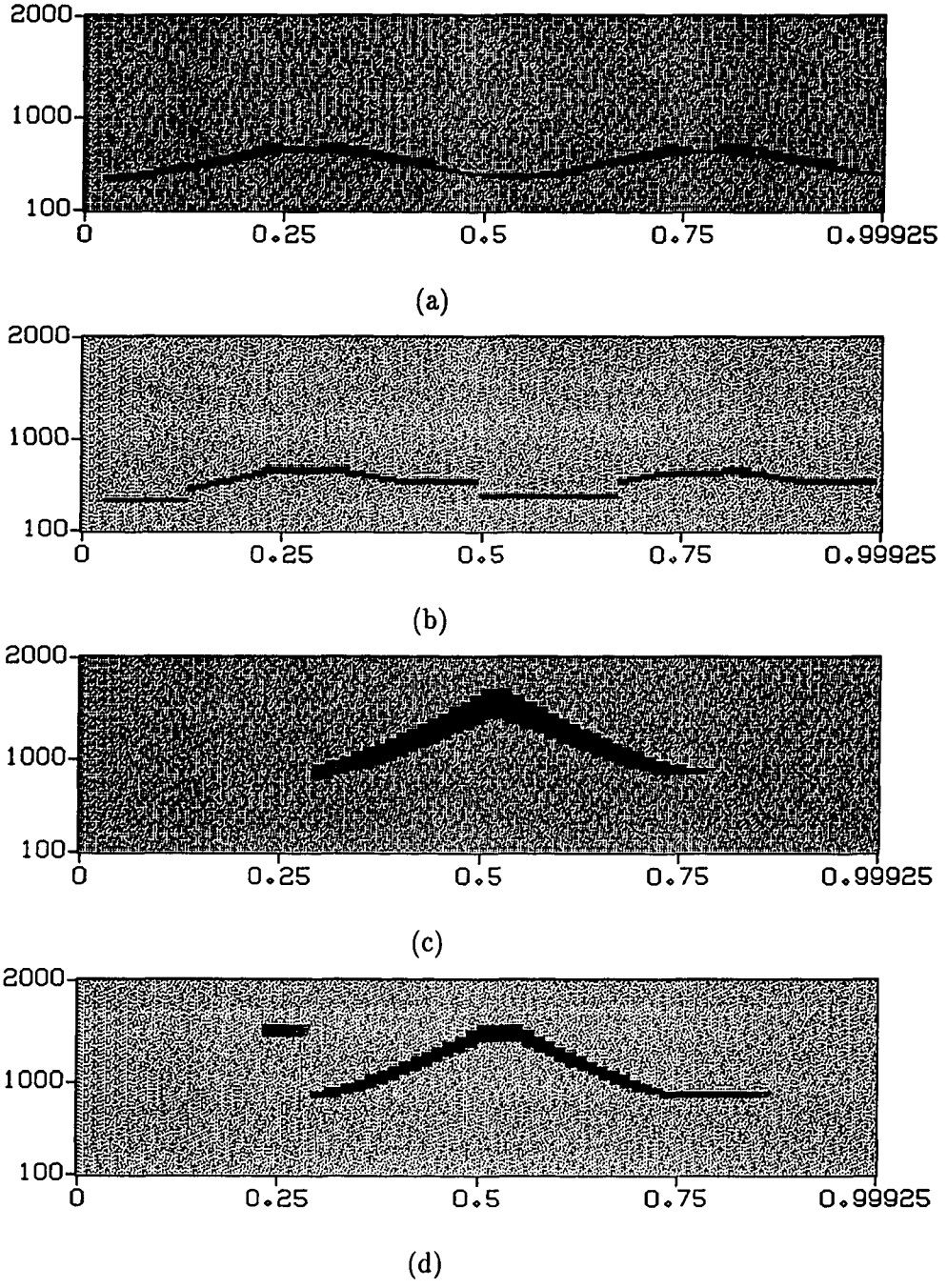


Figure 4·32: Model results for the Steiger (1980) diamond stimulus with noise bursts replacing the bifurcation points. (a) spectral stream layer and (b) pitch stream layer for stream 1; and (c) spectral stream layer and (d) pitch stream layer for stream 2.

4.6 Extension to model: Spatial location

This section outlines how spatial location cues can be incorporated into the model to aid the segregation process. The spatial location cues indirectly influence grouping by assisting grouping based on pitch. Thus, spatial cues by themselves cannot group objects, but require a pitch difference to exist, in keeping with the data from Shackleton, Meddis, and Hewitt (1994).

4.6.1 Spatial location cues

The auditory system localizes sounds using two different mechanisms: interaural time differences (ITD) and interaural intensity differences (IID). The concept behind both ITD and IID is that the listener is comparing the signal between the two ears (interaural) and making a judgment on the sound's location.

ITD, which operates at low frequencies (less than 5 kHz), corresponds to comparing the arrival time of a signal to the two ears. If a signal is to the left, it will arrive at the left ear some microseconds before it arrives at the right ear. Thus at 0 ITD, the source is centralized, and at other ITDs the source is more lateral. However, ITDs only work for low frequency, where the wavelength is long compared to the size of the head. Figure 4·33 shows a schematic representation of an object that is lateralized to the right. As the object emits a sound, it will arrive at the right ear first, and then at the left ear τ microseconds later, corresponding to the extra path distance d that the source has to travel.

At high frequencies, the head "shadows" a sound lateralized to one side, causing an IID, or intensity difference. For example, if a high frequency sound is located to the left, the intensity of the sound to the right ear is diminished compared to the left ear. Thus, one can localize the sound by some computation based on the intensity difference at the two ears. The extended model presented here incorporates only

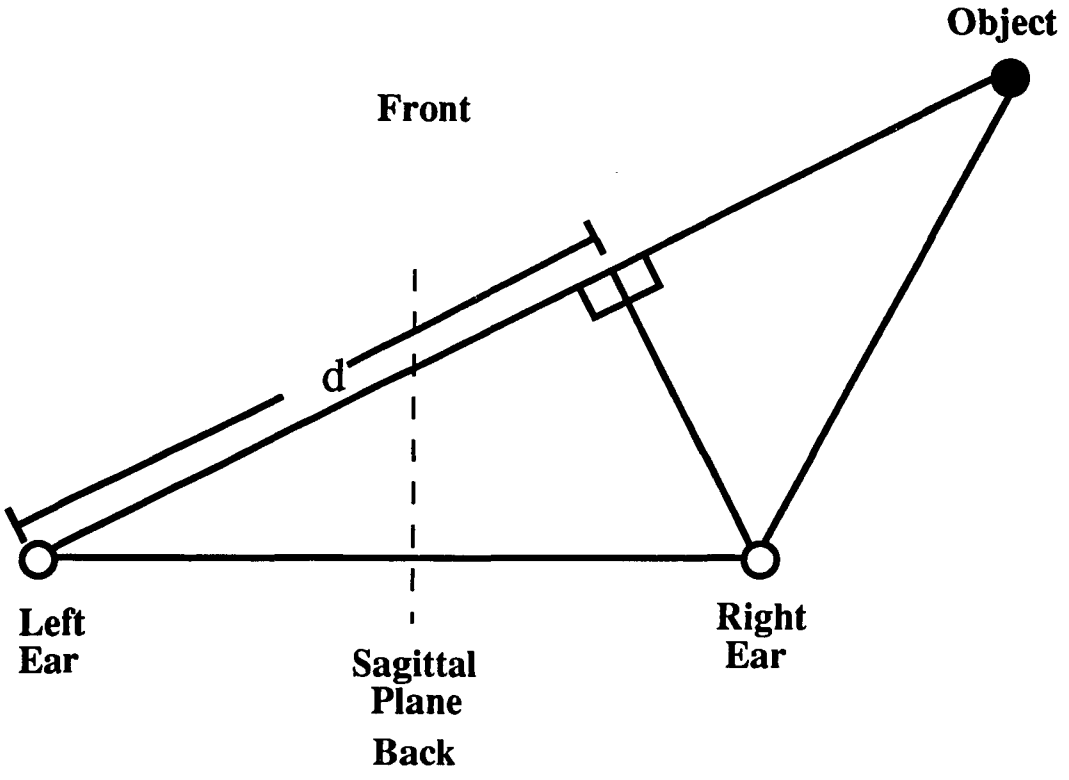


Figure 4·33: Geometric representation of spatial lateralization using interaural timing differences (ITD).

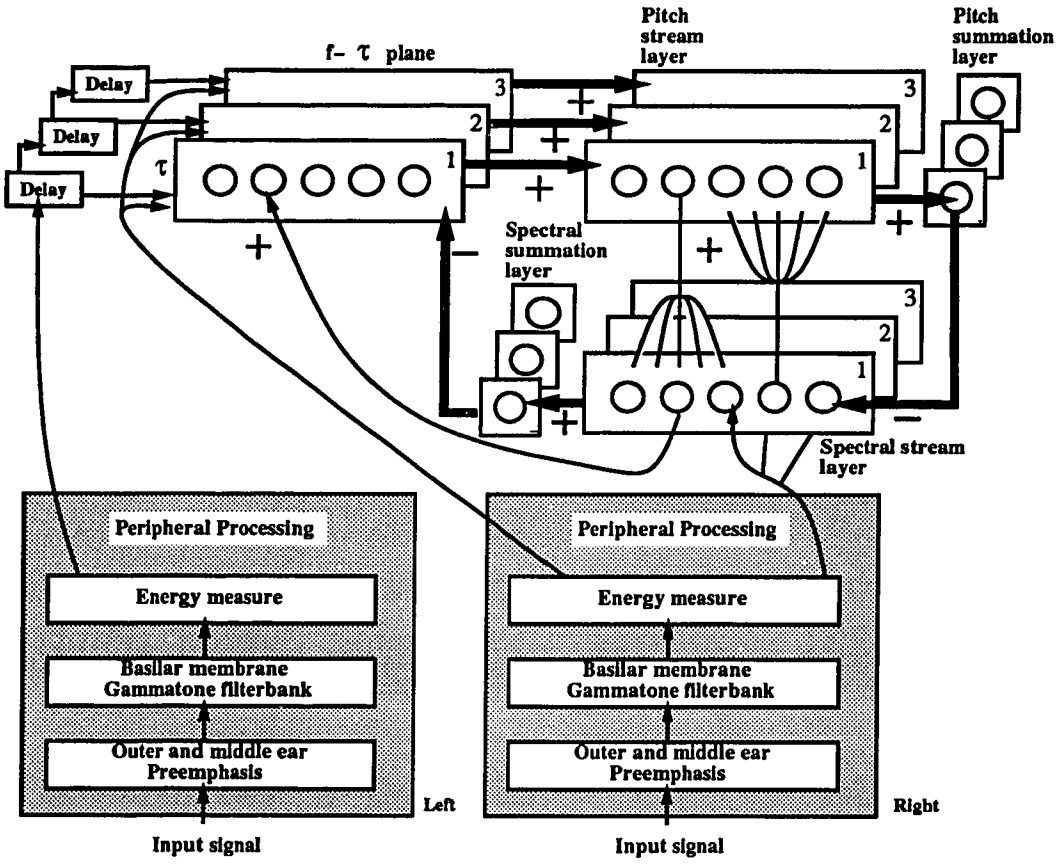


Figure 4·34: Block diagram of the extended streaming model.

ITDs in the segregation process.

4.6.2 Extended model

The extended model is shown in Figure 4·34. The model first preprocesses the incoming signal in the peripheral processing modules. This preprocessed signal is then used to determine spatial locations for the frequency components, and at the same time group frequency components based on pitch using the spectral and pitch stream layers from the original model. Segregation of components is accomplished in the pitch and spectral stream layers; the spatial locations non-specifically prime

their corresponding pitch stream layer to bias them towards grouping components. Next, those components which have been grouped by pitch are reinforced based on their spatial locations.

The peripheral preprocessing is identical for both the left and right "ears", and consist of the same module as in the original model. The output of this peripheral processing is fed to the f-τ plane, where individual frequencies f are assigned to a spatial location τ. τ represents radial direction, taking on values from -600 to 600 μs. The value $\tau = 0$ corresponds to the central location, which is a location centered between the "ears" and in front of the listener; $\tau = -600$ corresponds to a location that is directly to the left of the listener; and $\tau = 600$ corresponds to a location that is directly to the right of the listener. It is assumed that τ maps to radial direction in a linear fashion. It is also assumed that only one stream can occupy one spatial location, except at the central "head-centered" location, where multiple streams can be represented. Once components have been assigned to a given location, the location non-specifically primes all the neurons in its corresponding pitch stream layer. Figure 4·35 shows how the spatial locations non-specifically primes the pitch stream layers, and how a frequency component at a given spatial location in the f-τ is reinforced by its corresponding frequency component in the spectral stream layer.

The output of the right channel also feeds into the different streams of the spectral stream layer. The spectral stream layers are the same as in the original model. The pitch stream layer is modified so that all neurons within a stream get excited if there are any components present at that given location. Thus, a pitch stream layer will be biased to win over another pitch stream layer if there are components present at that location. At the central location, the N streams are all excited. In addition, the asymmetric competition across streams, term $L \sum_{k>i} g(P_{kp})$ in equation 4.20, exists only at the central location; non-central streams equally inhibit each other.

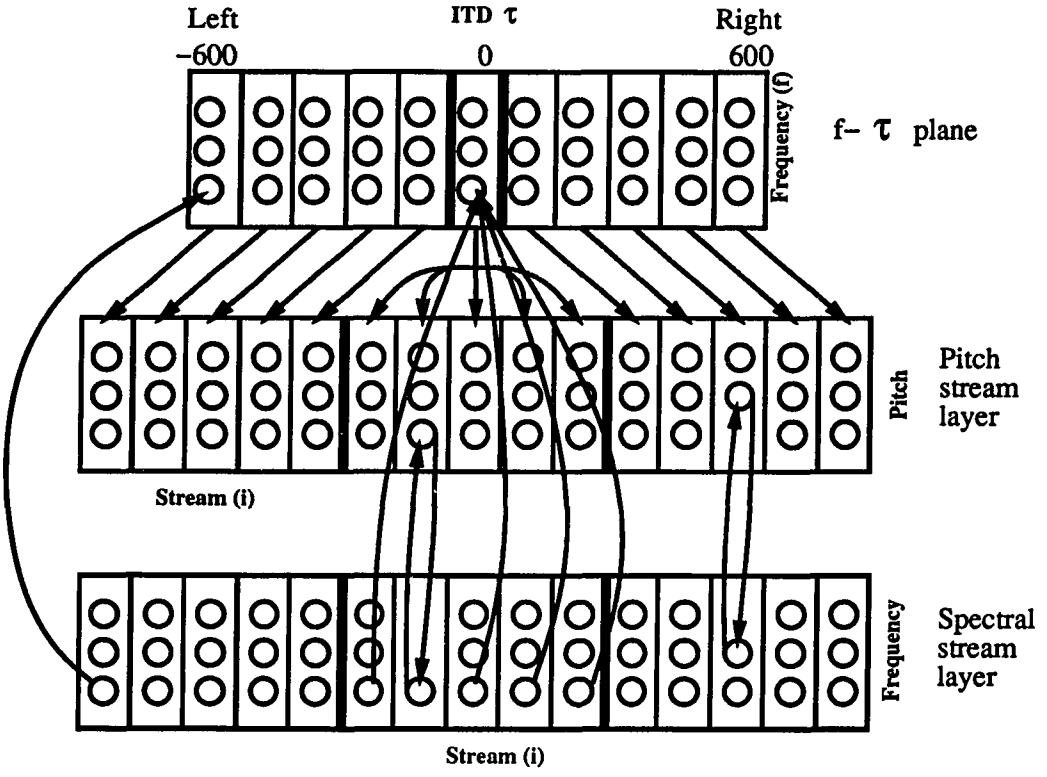


Figure 4-35: Interaction between spatial locations in the f-τ field, pitch stream layer, and the spectral stream layer. The non-specific inhibitory neurons are not shown. Only one stream can occupy one spatial location, except at the central "head-centered" location $\tau = 0$, where multiple streams can be represented. Once a spatial location has been derived for all the components, the spatial location non-specifically primes all the neurons in its corresponding pitch stream layer. At the central location, the N streams are all primed. Once components have been grouped based on pitch, the neurons in a spectral stream layer specifically excite the components at their corresponding spatial location. At the central location, the spectral neurons, corresponding to a given frequency, from all N streams excite the corresponding neuron at $\tau = 0$.

In addition, there is feedback from the spectral stream layer back to the f-τ plane. The feedback consists of a specific excitatory feedback and a non-specific inhibitory feedback, akin to the connectivity from the pitch stream layer to the spectral stream layer. The specific feedback excites those harmonic components existing at a given location where a pitch has been determined. At the central location, the spectral neurons, corresponding to a given frequency, from all N streams excite the corresponding neuron at $\tau = 0$. The spectral summation layer provides non-specific inhibitory feedback to suppress those (inharmonic) frequency components that do not belong to that pitch, allowing other spatial locations to capture that frequency component, and in turn, leading to complete resonance within the model.

The extended model is capable of replicating the Deutsch (1975) scale illusion (Figure 4-8), where a downward and an upward scale are being played at the same time, except that every other tone in a given scale is presented to the opposite ear. The result is that listeners group based on frequency proximity, and hear a bounce percept. In order to understand how the model produces this phenomenon, one needs to recall that the extended model does not group based on spatial location, but instead, spatial location only primes the grouping based on pitch process. For the first two simultaneous tones, hi C presented to the left ear and a low C presented to the right ear, the left and right spatial locations become active, priming their corresponding pitch stream layers. This in turn causes the left stream to capture the hi C tone and the right stream to capture the low C tone. For the next two simultaneous tones, a B presented to the right ear and a D presented to the left ear, both the left and right channels are still equally active, which causes both the left and right pitch stream layers to remain equally primed. Now, due to frequency proximity in the spectral stream layer, the B will be grouped with the hi C tone, and the D will be grouped with the low C tone. Thus, due to equal activation of the

left and right spatial locations, grouping based on frequency proximity overcomes grouping based on spatial location. Similarly, the rest of the tones in the sequence will be grouped based on proximity, leading to the bounce percept.

4.7 Summary

This chapter presented a model of auditory scene analysis that suggests how the brain segregates overlapping auditory components using pitch cues to create different mental objects. The model is shown to qualitatively replicate listeners' percepts of hearing two streams for two inharmonic tones, the continuity illusion, a bounce percept for crossing glides even if the intersection point is replaced by silence or noise, and the "M" and inverted "V" percept for the Steiger (1980) diamond stimulus even if the bifurcation points are replaced by noise. This chapter also presented how spatial cues can be incorporated into the model. While the extended model has been outlined, it needs to instantiated to verify that it is capable of producing correct percepts, such as the Deutsch (1975) scale illusion.

While the model is capable of qualitatively producing correct responses for the stimuli mentioned above, the model needs to incorporate other mechanisms in order to emulate other phenomena. The existing model does not contain any onset or offset cues to help create more veridical percepts, e.g. the spectral layer decays slowly at the offset of a tone. In addition, the onset/offset cues can influence the segregation process, e.g. the continuity illusion of hearing a tone in noise can be destroyed by decreasing or increasing the amplitude of the tone at the onset/offset of the noise. Another set of data that needs to be investigated consists of how the addition of harmonics can help overcome grouping by proximity, e.g. the addition of harmonics to one glide in a crossing glide stimulus leads to a cross percept and not a bounce percept. Finally, no learning exists in the model, and thus an exploration of how an

organism can learn to self-organize this substrate for auditory scene analysis needs to be explored.

References

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America, 51*, 648–651.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In Fant, G., & Tatham, M. A. A. (Eds.), *Auditory Analysis and Perception of Speech*, pp. 103–113. Academic, London.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America, 71*(4), 975–989.

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Toward an auditory theory of speaker normalization. *Lang. and Comm., 4*, 59–69.

Boardman, I., Cohen, M. A., & Grossberg, S. (1993). Variable rate working memories for phonetic categorization and invariant speech perception. In *Proceedings of the World Congress on Neural Networks*, Vol. III, pp. 2–5 Hillsdale, NJ. Erlbaum Associates.

Bregman, A. S., & Tougas, Y. (1989). Propagation of constraints in auditory organization. *Perception and Psychophsyics, 46*, 395–396.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Cambridge, MA.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psych., 89*, 244–249.

Bregman, A. S., & Dannenbring, G. (1973). The effect of continuity on auditory stream segregation. *Perception and Psychophsyics, 13*, 308–312.

Bregman, A. S., & Doehring, P. (1984). Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception and Psychophsyics, 36*, 251–256.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology, 32*, 19–31.

Bregman, A. S., & Rudnicky, A. (1975). Auditory segregation: Stream or streams?. *J. Exp. Psych.: Human Perception and Performance, 1*, 263–267.

Bregman, A. S., & Steiger, H. (1980). Auditory streaming and vertical localization: Interdependence of 'what' and 'where' desicions in audition. *Perception and Psychophsyics, 28*, 539–546.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America, 29*, 708–710.

Brokx, J. P. L., & Noteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. of Phonetics, 10*, 23–26.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America, 83*, 1508–1516.

Brown, G. J. (1992). *Computational auditory scene analysis: A representational approach.* Ph.D. thesis, University of Sheffield.

Carhart, R., Tillman, T. W., & Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *Journal of the Acoustical Society of America, 45*, 694–702.

Carlyon, R. P. (1991). Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America, 89*,

329–340.

Carlyon, R. P. (1992). The psychophysics of concurrent sound segregation. In Carlyon, R. P., Darwin, C. J., & Russell, I. J. (Eds.), *Processing of Complex Sounds by the Auditory System.* Clarendon Press, Oxford.

Carpenter, G. A., & Grossberg, S. (1991). *Pattern Recognition by self-organizing neural networks.* MIT Press, Boston, MA.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. Neural Networks*, *3*(5), 698–713.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991a). Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Carpenter, G. A., Grossberg, S., & Rosen, D. B. (1991b). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*(6), 759–771.

Chalika, M. H., & Bregman, A. S. (1989). The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Perception and Psychophsyics*, *46*, 487–497.

Cohen, M. A., Grossberg, S., & Wyse, L. (1992). A neural network for synthesizing the pitch of an acoustic source. Tech. rep. CAS/CNS-TR-92-009, Boston University, Boston, MA.

Cohen, M. A., Grossberg, S., & Wyse, L. (1994). A neural network spectral model of pitch detection and representation. Tech. rep., Boston University. (sumbitted).

Colburn, H. S. (1973). Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination. *Journal of the Acoustical Society of America, 54*, 1458–1470.

Colburn, H. S. (1977). Theory of binaural interaction based on auditory-nerve data. II. detection of tones in noise. *Journal of the Acoustical Society of America, 61*, 525–533.

Cooke, M. P. (1991). *Modelling auditory processing and organisation.* Ph.D. thesis, University of Sheffield.

Cutler, A., & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication, 9*, 485-489.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psych. Rev., 83*, 114–140.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex waves. *Perception and Psychophsyics, 24*, 369–376.

Darwin, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America, 76*, 1636–1647.

Darwin, C. J., & Bethell-Fox, C. E. (1977). Pitch continuity and speech source attribution. *J. of Exp. Psych: Human Perception and Performance, 3*, 665–672.

Darwin, C. J., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *Journal of the*

Acoustical Society of America, 91, 3381–3390.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic?. *Q. J. of Exp. Psych, 36A*, 193–208.

Dasarathy, B. V. (Ed.). (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, CA.

de Boer, E., & de Jongh, H. R. (1978). On cochlear encoding: Potentialities and limitations of the reverse correlation technique. *Journal of the Acoustical Society of America, 63*, 115–135.

Deschovitz, D. (1977). Information conveyed by vowels: A confirmation. *Haskins Lab status report on Speech Research, SR 51/52*, 213–219.

Deutsch, D. (1975). Two-channel listening to musical scales. *Journal of the Acoustical Society of America, 57*, 1156–1160.

Dorman, M. F., & Raphael, L. J. (1980). Distribution of acoustic cues for stop consonant place of articulation in VCV syllables. *Journal of the Acoustical Society of America, 67*(4), 1333–1335.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York.

Fant, G. (1966). A note on vocal tract size factors and non-uniform f-pattern scalings. *Q. Prog. Stat. Rep. STL-QPSR, 4/1966*, 22–30.

Fant, G. (1975). Non-uniform vowel normalization. *Q. Prog. Stat. Rep. STL-QPSR, 2-3/1975*, 1–19.

Gardner, R. B., Gaskill, S. A., & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America, 85*, 1329–1337.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63*, 223–230.

Gelfand, S. A., Ross, L., & Miller, S. (1988). Sentence perception in noise from one versus two sources: Effects of aging and hearing loss. *Journal of the Acoustical Society of America, 83*, 248–256.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Trans. on Audio and Electroacoustics, AU-16*(1), 78–80.

Grossberg, S. (1980). How does a brain build a cognitive code?. *Psychological Review, 87*, 1–51.

Hall, J. W., & Grose, J. H. (1988). Comodulation masking release: Evidence for multiple cues. *Journal of the Acoustical Society of America, 84*, 1669–1675.

Halpern, L. (1977). The effect of harmonic ratio relationships on auditory stream segregation. Tech. rep., McGill University, Psychology Department.

Hindle, D. (1978). Approaches to vowel normalization in the study of natural speech. In Sankoff, D. (Ed.), *Linguistic Variation: Models and Methods*, pp. 161–171. Academic, New York.

Jeffress, L. A. (1948). A place theory of sound localization. *J. Comp. Physiol. Psychol., 41*, 35–39.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in 'vowelless' syllables. *Perception and Psychophysics, 34*, 441–450.

Just, M. A., Suslick, R. L., Michaels, S., & Shockey, L. (1978). Acoustic cues and psychological processes in the perception of natural stop consonants. *Perception and Psychophysics, 24*(4), 327–336.

Klatt, D. H. (1992). Review of selected models of speech perception. In Marlsen-Wilson, W. D. (Ed.), *Lexical Representation and Process.* MIT Press, Cambridge.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*(1), 98–104.

Levitt, H., & Rabiner, L. R. (1967). Binarual release from masking for speech and gain in intelligibility. *Journal of the Acoustical Society of America, 42*, 601–608.

Liberman, A. M., Delattre, P. C., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychol. Mono., 68*, 1–13.

Macchi, M. J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America, 68*(6), 1636–1642.

McAdams, S. (1989). Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86*, 2148–2159.

Meddis, R., & Hewitt, M. J. (1992). Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America, 91*, 233–245.

Miller, G. A., & Licklider, J. C. R. (1950). Intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22*, 167–173.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73*, 1751–1755.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America, 77*, 1853–1860.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America, 74*(3), 750–753.

Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels.* Indiana University Linguistics Club, Bloomington, IN.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*(5), 2088–2113.

O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine.* Addison-Wesley, Reading, MA.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Port, R. F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, 7(1), 45–56.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *J. Exp. Psychol. Hum. Percept. Perf., 13*, 40–61.

Repp, B. H. (1980). A range-frequency effects on perception of silence in speech. *Haskins Lab. Status Report on Speech Research, SR-61*, 151–165.

Repp, B. H., Lieberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *JEP: Human Perception and Performance, 4*, 621–637.

Roberts, B., & Moore, B. C. J. (1991). The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony. *Journal of the Acoustical Society of America, 89*, 2922–2932.

Scheffers, M. T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation.* Ph.D. thesis, Groningen University.

Shackleton, T. M., Meddis, R., & Hewitt, M. J. (1992). Across frequency integration in a model of lateralization. *Journal of the Acoustical Society of America, 91*, 2276–2279.

Shackleton, T. M., Meddis, R., & Hewitt, M. J. (1994). The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels. Tech. rep., University of Technology. (submitted).

Sharf, D., & Ohde, R. (1984). Effects of formant frequency onset variation on the differentiation of synthesized /w/ and /r/ sounds. *Journal of Speech and Hearing Research, 27*, 475–379.

Steiger, H. (1980). Some informal observations concerning the perceptual organization of patterns containing frequency glides. Tech. rep., McGill University, Montreal.

Steiger, H., & Bregman, A. S. (1982). Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception and Psychophsyics, 32*, 152–162.

Strange, W., Verbrugge, R. R., Shankweiller, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, *60*(1), 213–224.

Summerfield, Q. (1992). Roles of harmonicity and coherent frequency modulation in auditory grouping. In Schouten, M. E. H. (Ed.), *Audition, Speech, and Language.* Mouton, Berlin.

Summerfield, Q., & Culling, J. F. (1992). Auditory segregation of competing voices: Absence of effects of FM or AM coherence. In Carlyon, R. P., Darwin, C. J., & Russell, I. J. (Eds.), *Processing of Complex Sounds by the Auditory System.* Clarendon Press, Oxford.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*(4), 1086–1100.

Tartter, V. C., Kat, D., Samuel, A. G., & Repp, B. H. (1983). Perception of intervocalic stop consonants: The contributions of closure duration and formant transitions. *Journal of the Acoustical Society of America*, *74*(3), 715–725.

Tougas, Y., & Bregman, A. S. (1990). The crossing of auditory streams. *J. of Exp. Psychol.: Human Perception and Performance*, *11*, 788–798.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, *69*(5), 1465–1475.

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences.* Ph.D. thesis, Eindhoven University of Technology.

Wakita, H. (1977). Normalization of vowels by vocal tract and its application to vowel identification. *IEEE Trans. ASSP*, *25*, 183–192.

Watrous, R. L. (1991). Current status of Peterson-Barney vowel formant data. *Journal of the Acoustical Society of America*, *89*(5), 2459–2460.

Watrous, R. L. (1993). Speaker normalization and adaptation using second-order connectionist networks. *IEEE Trans. Neural Networks*, *4*(1), 21–30.

Widrow, B., & Stearns, S. D. (1985). *Adaptive Signal Processing.* Prentice Hall, Englewood Cliffs, NJ.

Zahorian, S. A., & Jagharghi, A. J. (1991). Speaker normalization of static and dynamic vowel spectral features. *Journal of the Acoustical Society of America*, *90*(1), 67–75.

Zwicker, E., & Terhardt, E. (1980). Analytical expression for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, *68*, 1523–1525.

Curriculum Vitae

Krishna K. Govindarajan

EDUCATION

Boston University

PhD, Cognitive and Neural Systems, Sept. 1989 - Sept. 1994

University of Colorado, Boulder

BS, Electrical and Computer Engineering, Sept. 1984 - May 1989

WORK EXPERIENCE

Boston University, Research Fellow, Sept. 1990 - Aug. 1994

Research to develop and test neural network models of auditory and speech perception research with professors Gail Carpenter, Michael Cohen, and Stephen Grossberg. Projects include auditory scene analysis and source segregation, evaluation of speaker normalization methods, and adaptation effects in speech perception. Supported by grants from ARPA and NSF.

Boston University, Teaching Assistant, Jan. 1993 - May 1993

Aided in teaching and grading graduate course "Neural and Computational Models of Speech Perception and Production." Topics included physiology, psychophysics, and models of speech perception and production, including hidden markov models (HMM).

Draper Laboratory, Research Fellow, Sept. 1989 - Sept. 1990

Helped create a real-time neural network for heading and depth control of an autonomous submarine.

Precision Visuals, Inc., Engineer, Aug. 1988 - Aug. 1989

Assisted in development and release of scientific visualization software, PV-WAVE, for various platforms, including Sun, DEC, and Silicon Graphics.

National Center for Atmospheric Research, Student Assistant, Sept. 1986 - May 1987

Developed, maintained, and documented radar imagery software.

Professional Activities

Refereed manuscripts for *Neural Networks* journal and *World Congress on Neural Networks (1993)* conference. Organizer of temporal pattern recognition journal club, and member of auditory perception journal club. Member of IEEE and ASA.

Publications

1. Farrell, J., Goldenthal, B., and Govindarajan, K. K. (1990) "Connectionist learning control systems: Submarine depth control", Proceedings of the IEEE Conference on Decision and Control, Honolulu, Hawaii, 5-7 December 1990, pp. 2362-2367, IEEE Press, New York, NY. Technical Report CSDL-P-2982, Charles Stark Draper Laboratory, Inc., Cambridge, MA.

2. Carpenter, G. A., and Govindarajan, K. K. [1] (1993) "Speaker normalization methods for vowel recognition: Comparative analysis using neural network and nearest neighbor classifiers", submitted for publication. Technical Report CAS/CNS-TR-93-039, Boston University, Boston, MA.

3. Carpenter, G. A., and Govindarajan, K. K. [1](1993) "Evaluation of speaker normalization methods for vowel recognition using neural network and nearest neighbor

[1]Authors listed in alphabetical order.

classifiers", Journal of the Acoustical Society of America, **93**, pp. 2353. Oral presentation at the 125th Acoustical Society of America conference, Ottawa, Canada, 17-21 May 1993.

4. Carpenter, G. A., and Govindarajan, K. K. [1](1993) "Evaluation of speaker normalization methods for vowel recognition using fuzzy ARTMAP and K-NN", Proceedings of the World Congress on Neural Networks (WCNN-93), Portland, Oregon, 11-15 July 1993, pp. III-10–III-16, Lawrence Erlbaum Associates, Hillsdale, NJ. Oral presentation. Technical Report CAS/CNS-TR-93-013, Boston University, Boston, MA.

5. Govindarajan, K. K., and Cohen, M. A. (1994) "Influence of silence duration distribution in perception of stop consonant clusters", Journal of the Acoustical Society of America, **95**, pp. 2978. Poster presentation at the 127th Acoustical Society of America conference, Cambridge, MA, 6-10 June 1994.