

INTELLIGIBILITY OF FILTERED SPEECH AND ESTIMATION OF FREQUENCY-
IMPORTANCE FUNCTIONS

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Philip Loizou, Chair.

Dr. Aria Nosratinia

Dr. Murat Torlak

Copyright 2002

Kalyan S. Kasturi

All Rights Reserved

To my dear parents

INTELLIGIBILITY OF FILTERED SPEECH AND ESTIMATION OF FREQUENCY-
IMPORTANCE FUNCTIONS

by

KALYAN S. KASTURI, B.TECH.

THESIS

Presented to the faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

August 2002

ACKNOWLEDGEMENTS

First of all, I would like to express my deep sense of respect and gratitude towards my advisor Dr. Philip Loizou, who has been the guiding force behind this work. I want to thank him for introducing me to the field of Signal Processing and giving me the opportunity to work in various research projects. I am greatly indebted to him for his constant encouragement and invaluable advice in every aspect of my academic life. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

Next, I want to express my respects to Dr. Aria Nosratinia for teaching me random processes, but also helping me how to learn. He has been a great source of inspiration to me and I thank him for obliging to be on my defense committee and providing valuable suggestions.

I thank Dr. Murat Torlak for being kind to agree to serve on the committee and giving useful feedback on this manuscript.

I also express my respects and thanks to Dr. Mohammad Saquib for his encouragement and good advice. I also thank Dr. Fonseka, Dr. Arthur Lobo, Dr. Oguz Poroy and Lakshmi Mishra for their kind interest and help. Thanks are also due to all of my lab mates, from whom I learned a lot and whose companionship I enjoyed very much.

Lastly, I would like express my gratitude to NIDCD/NIH (Grant No.R01 DC03421) for their support.

INTELLIGIBILITY OF FILTERED SPEECH AND ESTIMATION OF FREQUENCY-
IMPORTANCE FUNCTIONS

Kalyan S. Kasturi, M.S.E.E.
The University of Texas at Dallas, 2002

Supervising Professor: Dr. Philip C. Loizou

An understanding of how information about the speech signal is spread among the various frequency bands of the spectrum is essential in numerous communications, audio and hearing related applications. Although many studies investigated the intelligibility of high-pass, low-pass and band-pass filtered speech, not many studies investigated the perception of band-stop filtered speech (i.e., speech with holes in the spectrum) or speech composed of disjoint frequency bands. The most recent studies examined speech recognition either for a single hole varying in frequency location and size or for a single hole in the middle of the spectrum. The scope of these studies is limited in the sense that they did not consider perception of speech composed of multiple disjoint bands involving low, middle and/or high frequency information. The present study addresses this question in a systematic fashion, considering all possible combinations of missing disjoint bands from the spectrum. In this work, we also derive frequency-importance functions for consonant and vowel recognition using (a) a least

squares approach that utilizes the results of intelligibility tests for speech with holes in the spectrum and (b) an information theoretic approach based on the calculation of mutual information between frequency bands and phonetic labels.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	vi
LIST OF FIGURES.....	x
LIST OF TABLES.....	xiv
1. INTRODUCTION.....	1
2. LITERATURE REVIEW.....	4
2.1 Chapter outline.....	4
2.2 Intelligibility of filtered speech.....	4
2.3 Frequency-importance functions.....	7
2.3.1 Articulation index approach.....	7
2.3.2 Correlation based approach.....	10
2.3.3 Information theoretic approach.....	16
3. INTELLIGIBILITY OF SPEECH WITH HOLES IN THE SPECTRUM.....	27
3.1 Chapter outline.....	27
3.2 Motivation.....	27
3.3 Method.....	30
3.3.1 Subjects.....	30
3.3.2 Speech material.....	31
3.3.3 Signal processing.....	31
3.3.4 Procedure.....	34
3.4 Results.....	36

3.4.1	Single-hole conditions.....	36
3.4.2	Two-hole conditions.....	38
3.5	Discussion.....	41
3.5.1	Effect of the location of spectral “holes”.....	42
3.5.2	Effect of size and pattern of spectral “holes”.....	45
4.	FREQUENCY IMPORTANCE FUNCTIONS.....	48
4.1	Chapter outline.....	48
4.2	Least squares approach.....	48
4.2.1	Limitations of articulation index.....	48
4.2.2	Estimation of weights.....	49
4.2.3	Results.....	51
4.2.4	Discussion.....	52
4.3	Information theoretic analysis.....	56
4.3.1	Processing of speech data.....	56
4.3.2	Computation of spectral energy.....	56
4.3.3	Quantization of spectral energy.....	57
4.3.4	Calculation of probability distributions.....	57
4.3.5	Computation of joint distributions.....	59
4.3.6	Calculation of mutual information.....	59
4.3.7	A summary of implementation procedure.....	60
4.3.8	Results and discussion.....	64
5.	SUMMARY AND CONCLUSIONS.....	71
	APPENDIX A.....	74
	APPENDIX B.....	79
	REFERENCES.....	87

VITA

LIST OF FIGURES

Figure 3.1: Block diagram representing the signal processing performed.....	33
Figure 3.2: Mean percent scores for vowel and consonant recognition as a function of the location of the spectral “hole”. The “holes” were centered around the channel center frequencies. In the baseline condition, all channels were present.....	36
Figure 3.3: Percent information transmitted for the features place, manner and voicing as a function of the location of the spectral “hole”.....	37
Figure 3.4: Mean percent scores for vowel and consonant recognition as a function of the location of the pair of spectral “holes”. The “holes” were introduced at frequencies centered at the channel pairs indicated. In condition (1,4), for instance, channels 1 and 4 were removed from the spectrum. In the baseline condition, all channels were present....	39
Figure 3.5: Percent information transmitted for the features place, manner and voicing as a function of the location of the pair of frequency bands removed.....	40
Figure 3.6: Mean percent scores on individual vowel recognition for the condition in which channel 2 was removed from the spectrum (n=20). The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively. Error bars indicate standard errors of the mean.....	43
Figure 4.1: Frequency-importance function for consonants.....	51
Figure 4.1: Frequency-importance function for vowels.....	51
Figure 4.3: Individual listener’s frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for two subjects with the highest vowel scores, panels (c) and (d) show the functions for two subjects with middle scores, and panels (e) and (f) show the functions for two subjects with the lowest vowel scores.....	53
Figure 4.4: Quantized frame energy plots versus frame energy plots.....	61

Figure 4.5: Energy distribution for channels 1, 2 and 3.....	62
Figure 4.6: Energy distribution for channels 4, 5 and 6.....	63
Figure 4.7: Mutual information between frequency bands and phonetic labels for the syllable /apa/.....	64
Figure 4.8: Mutual information between the spectral energy and the phonetic labels for consonant stimuli.....	65
Figure 4.9: Mutual information between the spectral energy and the phonetic labels for vowel stimuli by female speakers.....	66
Figure 4.10: Mutual information between the spectral energy and the phonetic labels for vowel stimuli by male speakers.....	67
Figure 4.11: Comparison of weights obtained from mutual information and least squares methods for consonant stimuli.....	68
Figure 4.12: Comparison of weights obtained from mutual information and least squares methods for vowel stimuli by female speakers.....	69
Figure 4.13: Comparison of weights obtained from mutual information and least squares methods for vowel stimuli by male speakers.....	69
Figure A.1: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 1 and 2 respectively, panels (c) and (d) show the functions for subjects 3 and 4 respectively.....	74
Figure A.2: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 5 and 6 respectively, panels (c) and (d) show the functions for subjects 7 and 8 respectively.....	75
Figure A.3: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 9 and 10 respectively, panels (c) and (d) show the functions for subjects 11 and 12 respectively.....	76

Figure A.4: Individual listener’s frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 13 and 14 respectively, panels (c) and (d) show the functions for subjects 15 And 16 respectively.....	77
Figure A.5: Individual listener’s frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 17 and 18 respectively, panels (c) and (d) show the functions for subjects 19 And 20 respectively.....	78
Figure B.1: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 0 and 1 respectively, panels (c) and (d) show mean percent correct scores for the conditions 2 and 3 respectively. Error bars indicate standard errors of the mean.....	79
Figure B.2: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively, panels (c) and (d) show mean percent correct scores for the conditions 6 and 7 respectively. Error bars indicate standard errors of the mean.....	80
Figure B.3: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 9 and 15 respectively, panels (c) and (d) show mean percent correct scores for the conditions 16 and 18 respectively. Error bars indicate standard errors of the mean.....	81
Figure B.4: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively. Error bars indicate standard errors of the mean.....	82
Figure B.5: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 0 and 1 respectively, panels (c) and (d) show mean percent correct scores for the conditions 2 and 3 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively.....	83

Figure B.6: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively, panels (c) and (d) show mean percent correct scores for the conditions 6 and 7 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively..... 84

Figure B.7: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 9 and 15 respectively, panels (c) and (d) show mean percent correct scores for the conditions 16 and 18 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively..... 85

Figure B.8: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively..... 86

LIST OF TABLES

Table 3.1: The first two formant frequencies (in Hz) of the male and female vowels used in this study.....	31
Table 3.2: The 3-dB frequency boundaries of the 6 bands with the corresponding center frequencies (Hz) of each band.....	32
Table 3.3: The 22 test conditions considered in this study. The 0-th condition corresponds to the baseline condition.....	35
Table 4.1: Phonemes in American English.....	58

CHAPTER ONE

INTRODUCTION

The topic of speech perception or speech intelligibility has received major attention from researchers in speech and audio processing for several decades owing to its prime importance in the field of speech communication. Speech signal is a series of sounds rapidly varying from instant to instant in frequency and intensity. An understanding of how information about speech signal is spread among the various frequency bands of the spectrum is essential in numerous communications, audio and hearing related applications. The early research in this field was initiated at Bell Labs in the early years of 20th century to quantify the bandwidth of telephone line for speech communication.

Naturally, early research examined the intelligibility of speech under the conditions of high-pass and low-pass filtering conditions. Later studies included the investigation of intelligibility of band-pass filtered speech. But not many studies investigated the perception of band-stop filtered speech (i.e., speech with holes in the spectrum) or speech composed of disjoint frequency bands.

The most recent studies examined speech recognition either for a single hole, varying in frequency location and size, or for a single hole in the middle of the spectrum. The scope of these studies was limited in the sense that they did not consider perception of speech composed of multiple disjoint bands involving low, middle and/or high frequency information. The present study addresses this question in a systematic fashion considering all possible combinations of missing disjoint bands from the spectrum.

The answer to the question of how listeners use and combine information across frequency regions, whether isolated or disparate, is not only important for understanding speech perception in general but it is also important for understanding speech perception by cochlear implant listeners. Cochlear implants are based on the idea that there are surviving neurons in the vicinity of the electrodes. The lack of hair cells and/or surviving neurons in certain areas of the cochlea essentially creates hole(s) in the spectrum. The extent of the effect of holes in the spectrum on speech understanding is not well understood.

The aim of this study is to examine the effect of the location and size of spectral holes on vowel and consonant recognition. Understanding this effect will provide us with some insights as to why some cochlear implant listeners are not performing well, despite the wealth of information they receive.

In addition, we could use the data of this study to derive frequency importance functions that take into account the fact that listeners could combine information from disparate frequency regions in the spectrum. Several investigators have used the Articulation Index (AI) method to determine frequency importance functions. The AI method assumes that the information contained in each band is independent of the information contained in other bands and does not take into account the fact that listeners may combine speech information from multiple disjoint bands.

In this study, a frequency-importance function based on a least squares approach is proposed. Unlike the AI method, the proposed least squares approach makes use of the listeners' scores on perception of vowels and consonants composed of disjoint frequency bands. An alternative method for the estimation of frequency-importance functions based on

the calculation of mutual information between frequency bands and phonetic labels is also investigated.

This thesis is organized as follows; Chapter 2 presents a review of related literature in the field of speech intelligibility. In Chapter 3, current work involving the intelligibility of speech with holes in the spectrum and the details of intelligibility tests is presented. In Chapter 4, frequency-importance functions for consonant and vowel recognition are derived using (a) a least squares approach and (b) an information theoretic approach based on the calculation of mutual information between frequency bands and phonetic labels. Finally in Chapter 5 we summarize the work performed and present the conclusions.

CHAPTER TWO

LITERATURE REVIEW

2.1 Chapter Outline

It is an intriguing question as to how information pertaining to speech intelligibility is distributed among various frequency bands of speech. This chapter presents a review of related literature beginning with the general approaches used to answer the above question.

Literature dealing with the intelligibility of speech under different conditions of filtering is discussed in Section 2.2. In particular, studies regarding intelligibility of low-pass, high-pass and band-pass filtered speech are discussed.

Literature dealing with frequency importance functions is presented in Section 2.3. In particular, three different approaches based on the articulation index, correlation and information theory are discussed.

2.2 Intelligibility of filtered speech

2.2.1 Intelligibility of low-pass and high-pass filtered speech

French and Stienberg [16] investigated the intelligibility of low-pass and high-pass filtered speech. The test material consisted of meaningless monosyllables of the consonant-vowel-consonant type. The intensity and frequency content of the speech material was manipulated by the use of attenuators and filters. Intelligibility was measured as the percentage of syllables in which all the component sounds were perceived correctly. They reported an

average intelligibility of 90% for nonsense syllables low-pass filtered at 3.3 kHz and a relatively low intelligibility of 30% for the same stimuli high-pass filtered at 2.9 kHz.

Pollack [33] also studied the intelligibility of words where the speech material was subjected to the conditions of low-pass and high-pass filtering. Lists of monosyllabic words subjected to various filtering conditions were read by two trained talkers to nine experienced listeners. He reported a high intelligibility of 90% for speech low-pass filtered at 3.95 kHz, where as the intelligibility fell to 5% when speech was high-pass filtered at 2.375 kHz.

2.2.2 Intelligibility of band-pass filtered speech

Warren *et al.* [43] investigated the intelligibility of band-pass filtered speech with very narrow bands. *CID* sentences (“everyday speech”) and monosyllabic words were filtered with center frequencies of 370, 530, 750, 1100, 1500, 2100, 3000, 4200 or 6000 Hz. An average intelligibility greater than 95% was obtained for 1/3-octave bands centered around 1100, 1500 and 2100 Hz for sentences, while an average intelligibility greater than 50% was obtained for words. On the other hand, a relatively low intelligibility of around 24% was obtained with center frequencies 370 and 6000 Hz.

2.2.3 Intelligibility of speech with a single hole

Lippmann [25] investigated the intelligibility of consonants with a single hole in the middle of the spectrum. The speech material consisted of eight lists of consonant-vowel-consonant nonsense syllables. The speech material was processed through a combination of a low-pass filter and a high pass filter. The low-pass filter cutoff frequency was always set to 800 Hz. The high-pass filter cutoff frequency was varied between 3.15, 4, 5, 6.3, 8 and 10 kHz. High

consonant recognition was maintained even after removing speech energy in the middle frequencies (800 to 4 kHz).

Synergy effects were demonstrated in the study by Riener *et al.* [35] when subjects were presented with spectral information contained in the low and high frequency bands. The intelligibility of sentences through single one-third octave bands centered around 370 Hz and 6000 Hz was roughly 24% when presented alone, but increased to 77% when presented simultaneously.

Breeuwer and plomp [7] investigated intelligibility of speech using speechreading supplemented with frequency-selective sound-pressure information. They used short sentences to test eighteen normal hearing subjects. The intelligibility of sentences through one octave band centered around 500 Hz was 65.7%, but increased to 86.7% in the presence of an additional one octave band centered around 3160 Hz.

Shannon *et al.* [37] assessed the impact of the size and location of spectral holes with cochlear implant and normal hearing listeners. For the normal-hearing listeners, holes were created by dropping off 2 to 8 low, middle or high-frequency bands in a 20 noise-band cochlear implant (*CI*) simulation. Thus a single hole with varying size was introduced in the low, middle or high frequency regions of the spectrum. Results showed that holes in the low frequency region were more damaging than the holes in the middle and high frequency regions on speech recognition.

The current work focuses on the intelligibility of speech with multiple holes in the spectrum and investigates how information is combined across multiple disjoint bands to understand speech.

2.3 Frequency Importance Functions

2.3.1 Articulation index approach

Fletcher [15] developed the concept of articulation index while working towards understanding the basic principles behind human speech recognition. Fletcher used the term “articulation” as the probability of correct recognition for nonsense words (sounds having no meaning). Articulation index can be thought of as a speech recognition measure that accurately characterizes speech intelligibility under the conditions of filtering and noise.

A detailed discussion of the articulation index approach was given by French and Steinberg [16]. They performed articulation tests under different conditions of filtering and noise masking. The main aim of their study was to relate the articulation or percent correct recognition score of each test condition to a base value. In other words, the articulation score corresponding to a particular test condition could be expressed by an equivalent articulation index value. The articulation index values were expressed as a fraction of the maximum articulation index value corresponding to the test condition spanning the entire frequency band under optimum listening conditions. The maximum articulation index value was also referred to as the total articulation index.

Articulation index was used to establish relationships between the intelligence carrying capacity of the components of speech and their frequency. The speech was divided into n frequency bands. Each frequency band was associated with an articulation index value, which denoted the contribution of that band to the total articulation index. The articulation index of a particular frequency band was denoted by ΔA_i . It varied from zero to a maximum value A_0 and was given by:

$$\Delta A_i = W_i \cdot A_0 \quad (2.1)$$

Where A_0 was the articulation index corresponding to the entire frequency band under optimum listening conditions and W_i was a weight associated with the particular band of frequency for a particular value of signal-to-noise ratio.

The articulation indices of individual bands would all sum to the articulation index of the entire frequency band as given by:

$$A_0 = \sum_{i=1}^n \Delta A_i = \sum_{i=1}^n W_i \cdot A_0 \quad (2.2)$$

The concept of articulation index is based on the assumption that each band of frequency contributes to the total articulation index and its contribution is independent of the contributions of other frequency bands. The weights W_i 's were obtained by performing articulation tests on various low-pass and high-pass filtered versions of the speech database. The procedure employed distortion-less attenuators and amplifiers for varying the absolute level of speech.

One of the important frequency importance functions used for calculation of articulation index was reported in the *ANSI S3.5-1969* standard [3]. The function was strongly unimodal, with pronounced importance in the frequency range 1600-3150 Hz. The crossover frequency was 1660 Hz. This function was based on the data collected by the Bell Telephone Laboratories that corresponded to the work by Steinberg [16]. A problem associated with that function was that it was based on articulation tests for nonsense syllables alone.

Earlier investigations [15] [16] used only nonsense syllables tests for calculating the frequency importance functions in the computation of the articulation index. Later studies [12] [30] [40] indicated that the phonemic composition and format of the test material affected the calculation of the frequency importance functions in the computation of the articulation index.

Duggirala *et al.* [12] studied the effect of different phonemic features on frequency importance while fixing other influencing factors like message redundancy, talker characteristics etc. They used the articulation index approach developed by Steinberg [16] to estimate the relative importance of different frequency bands of the speech spectrum for the recognition of the Diagnostic Rhyme Test (*DRT*) and its six speech feature subtests.

In their experiment, eight scramblings of form *IV* of the *DRT* spoken by a male talker (*RH*) were used to test the subjects. This test was referred to as *DRT-IV-RH*. The stimulus words were corrupted with speech shaped noise to observe changes in performance by varying the signal level. The speech test consisted of 70 listening conditions corresponding to 14 filters and 5 signal levels. The articulation scores for each feature were averaged across subjects and trials and plotted as a function of filter cutoff frequencies.

The importance functions were computed using the collected data as mentioned above according to the graphical techniques described by Steinberg [16]. They found out that frequency importance function for the entire *DRT-IV-RH* test was more uniform across frequency than that developed by Steinberg [16] for nonsense syllable test. *DRT-IV-RH* reflects relatively greater weight in the frequency range 400-800 Hz and relatively less weight in the frequency range 1600-5000 Hz.

More importantly, the results of the six subtests of *DRT-IV-RH* indicated that the phonemic content of the test material greatly influenced the importance function. The subtests with phonemic content suitable for the evaluation of the speech features nasality and voicing resulted in importance functions located entirely below 1500 Hz. The subtests for the evaluation of the place of articulation features, graveness and compactness resulted in importance functions that were located mainly in the frequency range 500-4000 Hz. The subtests for the evaluation of the manner of articulation features sustention and sibilation resulted in importance functions that were located mainly in the frequency range 1600-8912 Hz.

2.3.2 Correlation-based approach

In correlation methods, weights are estimated according to a least-squares criterion [23]. The main advantage of correlation methods is that they can be generally applied in situations where statistical properties of the elements are unknown and where correlations exist between the elements.

Ahumada *et al.* [1] used multiple regression analysis to estimate the responses of the observer to signal tone plus noise stimulus as a linear combination of squared amplitudes of different frequency components. The main aim of their study was to estimate the features of the auditory stimulus that an observer uses to identify a signal tone masked by gaussian noise. The masking noise was represented as linear combination of sine waves having independent Raleigh-distributed amplitudes A_i , given by:

$$V_N(t) = \begin{cases} \sum_{i=0}^n A_i \sin(2\pi f_i t + \phi), & 0 \leq t \leq T \\ 0, & \text{elsewhere} \end{cases} \quad (2.3)$$

The signal tone (at frequency f_j) masked by noise was given by:

$$V_{SN}(t) = \begin{cases} V_N(t) + s \sin(2\pi f_j t), & 0 \leq t \leq T \\ 0, & \text{elsewhere} \end{cases} \quad (2.4)$$

Observers were presented with 16 sequences of 200 stimuli, of which 100 contained signal tones. The stimuli had 31 components from 350 to 650 Hz and varied in the amplitude distribution. The signal tone was 500 Hz. The response totals for each stimulus were computed for all the observers. Multiple regression analysis was used to obtain the coefficients c_i and intercept constants b that minimized the squared error of prediction P defined by:

$$P = \sum_{k=1}^m [R_k - (\sum c_i A_{ik}^2 + b)]^2 \quad (2.5)$$

where m was the number of stimuli used, R_k was the response total to k^{th} stimulus, A_{ik} was the amplitude of the i^{th} frequency component for the k^{th} stimulus. The linear regression coefficients were plotted versus different frequency components to obtain the observer's frequency response and it was observed that most of the subjects showed a peak at the tone frequency.

Richards and Zhu [34] presented a classical application of correlation analysis to the problem of signal detection. The signal detection model was based on the following assumptions: (a) independent one-dimensional decision variables are linearly combined to

form an ultimate decision variable, (b) final decision variable is then compared against a fixed criterion to form a decision variable. The decision variable was given by:

$$Z = \sum_{i=1}^n \alpha_i X_i \quad (2.6)$$

where Z was the decision variable, α_i were the combination weights and X_i were the independent random variables that formed subjective decision variables. The response variable T was given by:

$$T = \begin{cases} 1, & Z > C \\ 0, & \text{elsewhere} \end{cases} \quad (2.7)$$

where C is a fixed criterion. They assumed that the squared correlation between each X_i and T was proportional to α_i . For practical applications the variables like intensity or frequency are associated with the subjective variables X_i . For the case when internal noise in system was considered, it was assumed that noise was additive and normally distributed. A normal deviate was added to the decision variable Z , for the computation of relative combination weights. Computer simulations were performed to validate their propositions.

The paper by Lutfi [28] also gave a good discussion regarding the theory of the correlation approach, in particular the discrimination task where the observer received a sample of observations of random variables $x = \langle x_1, x_2, \dots, x_n \rangle$ that represented different elements of the stimulus. On each trial the observer had to decide between the two hypotheses:

$$H_0 : \sum a_j x_j \leq 0 \quad (2.8)$$

$$H_1 : \sum a_j x_j > 0 \quad (2.9)$$

where the coefficients a_j defined a specific task. The decision rule assumed was:

$$R = \begin{cases} R_1, & \text{if and only if } \sum_{j=1}^n w_j x_j + \varepsilon > C \\ R_0, & \text{otherwise} \end{cases} \quad (2.10)$$

where w_j were the observer's weights, ε was the observation error and C was a criterion.

The decision rule was given in a compact form as:

$$R = \sum_{j=1}^n w_j x_j - C + \xi \quad (2.11)$$

where $\xi \geq \varepsilon$ was the total error. If x_j were statistically independent, the above equation was found to reduce to the standard model for the analysis of multiple regression [10]. The weights were estimated according to the formula:

$$w_j = r_{Rx_j} (\sigma_R / \sigma_j) \quad (2.12)$$

where r_{Rx_j} was the point-biserial correlation observer's binary response and the value of the j^{th} stimulus element taken across trials.

Doherty and Turner [9] used a correlation-based method to determine a listener's weighting function for speech. Their method involved perturbing the various frequency components, and observing the effect of the amount of perturbation on the correct recognition of speech. The stimuli consisted of three vowel-consonant-vowel (VCV) signals. Digital linear phase band-pass filters were used to filter each of these signals into three separate frequency bands: (1) 200-750 Hz, (2) 750-2000 Hz and (3) 2500-9000 Hz. A

speech spectrum noise signal [12] was also filtered into the above three bands and each filtered noise band was then added to the corresponding filtered speech band at various SNR levels, over a 10-dB range in 2-dB steps (-7, -5, -3, -1, 1).

During each trial one of the five SNR levels was randomly selected for each band and a combination of all three bands were played to the listener. For each trial, listener's response, stimulus played and SNR levels used for each frequency band were recorded. For each listener, a point bi-serial correlation between listener's response and the SNR level within each band was computed. The correlations were then normalized to sum to one and plotted as relative weights. This plot was referred to as the weighting function of the listener.

Weighting functions were calculated for 6 subjects, where each weighting function consisted of three weights corresponding to the three bands. A large weight for a band indicated that the particular band was more important for speech recognition. The weights were found to vary across different subjects indicating that different subjects used different strategies for identifying speech. However, most of the subjects were found to place more weight on the second frequency band (750-2000 Hz). Also the same subjects were tested again and it was found that the weights did not vary much indicating the robustness of the correlation method.

Mehr *et al.* [29] used the correlation method to calculate the relative weights used for individual channels by multi-channel cochlear implant users. They calculated weighting functions for both normal hearing listeners and cochlear implant users to observe how the cochlear implant users differed from normal hearing listeners in the way they used the various frequency channels. Five normal hearing subjects and seven subjects using six-channel cochlear implants were tested. The test material consisted of a subset of University

of California, Los Angeles (*UCLA*) version of the Nonsense Syllable Test (*NST*) that uses six lists of consonant-vowel syllables.

They performed two experiments. In the first experiment, on each trial, a speech file was divided into six frequency bands corresponding to the six channels of the cochlear implant. By adding a random level of corresponding band-pass noise, each band was degraded. A combination of these noisy versions of all the bands was played to the subject. The subject's response and SNR values used for different bands were recorded. Around 1200 trials were performed and a rank order correlation between SNR values for different bands and subject's responses was computed, to obtain six correlation values corresponding to each band of cochlear implant. The correlations were normalized to sum to one to obtain the relative weights.

It was observed that in the case of normal hearing listeners, the relative weights were almost equal across different frequency bands. But in the case of cochlear implant users, the relative weights varied very much across different frequency bands with nearly zero relative correlation for at least one band. Thus in case of cochlear implant users some bands (specific to a particular subject) did not contribute to speech recognition.

In the second experiment, on each trial a particular band was removed and speech recognition was tested in quiet. The relative weights were computed as in the first experiment. For every subject it was observed that if a channel with a high relative correlation obtained from first experiment was removed, the speech recognition score dropped significantly. Thus the weights computed by correlation method were shown to be valid estimates.

2.3.3 Information theoretic approach

First we start with a historical perspective of information theory and cite some applications of information theory. We then proceed with some definitions and concepts that are central to information theory. Next we relate these concepts to the present work. We discuss the application of the concepts of mutual information and joint mutual information to study how the information pertaining to speech recognition is distributed across the spectrum.

Origin of Information Theory

The birth of information theory can be attributed to the revolutionary work of C.E. Shannon entitled “A Mathematical Theory of Communication” [36].

Information theory has found numerous applications that span many fields of science. Applications of information theory can be found in fields of engineering communications, computation, game theory, thermodynamics, statistical mechanics and many other fields. Even though information theory is ever pervasive, we shall confine our discussion to the field of data communications and signal processing that reaped the major benefits of information theory. This can be attributed to the fact that Shannon himself was especially concerned to push the applications to engineering communications. The modern communication theory owes a great debt to Shannon who showed that information transmission rate could be kept constant for an arbitrarily small probability of error. Earlier it was believed that information transmission rate over a noisy channel had to decline to zero for the error probability to approach zero. He used statistical models for communication channels and information sources to formulate the problem of robust transmission of information. More importantly Shannon established basic limits on communication of

information in terms of channel capacity. More specifically he demonstrated that reliable communication is possible only if the information rate from the source is less than the channel capacity.

Concept of Entropy

Entropy is the pivotal concept in information theory. Entropy of a random variable is a function of its probability distribution and is a good measure of uncertainty or randomness associated with the random variable. We shall confine our discussion to discrete random variables since they are more useful for modeling practical applications.

Consider a discrete random variable X with sample space $\langle x_1, x_2, \dots, x_n \rangle$. Let the probability measure be defined by $P(x_n) = p_n$. The entropy of X is defined as:

$$H(X) = \sum_{n=1}^N p_n \cdot \log(p_n) \quad (2.13)$$

The base of the logarithm is arbitrary. If the base is 2, the units of entropy are bits and if the base is e , the units are *nats* (natural units). $H(X)$ is a measure of uncertainty associated with the random variable X .

Entropy for Random Vectors

The concept of entropy can be extended to deal with random vectors as well. A random vector consists of a sequence of random variables. Consider a random vector $\bar{X} = (X, Y)$, where X and Y are random variables whose joint distribution is given by $p(x_i, y_j)$. Let N and

M be the number of sample points of random variables X and Y respectively. Then the entropy of \bar{X} is given as:

$$H(\bar{X}) = H(X, Y) = - \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log(p(x_i, y_j)) \quad (2.14)$$

An important and useful property of entropy that relates the entropy of a random vector to the entropy of individual random variables is given by:

$$H(X, Y) = H(Y) + H(X | Y) = H(X) + H(Y | X) \quad (2.15)$$

$H(X | Y)$ and $H(Y | X)$ are called the conditional entropies. $H(X | Y)$ is the conditional entropy of X given Y.

Mutual Information and its properties

$H(X | Y)$ gives the uncertainty associated with X, when information about Y is known. Then the quantity $H(X) - H(X | Y)$ gives the amount of reduction in uncertainty of X when Y is given. This quantity is defined to be the mutual information between the two random variables X and Y. It is denoted by $I(X, Y)$.

$$I(X, Y) = H(X) - H(X | Y) \quad (2.16)$$

Naturally the question arises how much information about Y is contained in X. According to the above definition of mutual information, we have $I(Y, X) = H(Y) - H(Y | X)$, where $I(Y, X)$ gives the amount of information about Y contained in X.

Mutual information satisfies the commutative property, that is $I(X, Y) = I(Y, X)$. This directly follows from the definitions of $I(X, Y)$ and $I(Y, X)$ and the Equation (2.15).

This is a very interesting result in the sense that X contains the same amount of information about Y , as does Y about X . Thus $I(X,Y)$ can be thought of as the average amount of information about X that is contained in Y and $I(X,Y)$ is called average mutual information between X and Y .

Another property of mutual information $I(X,Y)$ worth noting is that it is always non-negative. This follows from the definition of $I(X,Y)$ when used together with the well-known theorem that $H(X/Y) \leq H(X)$ where the equality holds if and only if X and Y are independent [18]. This also leads to the important observation that mutual information between two independent random variables is zero, which is consistent with our intuition.

Applications of Mutual Information

The mutual information between two random variables gives the average amount of information contained in one random variable about the other. A large value of mutual information indicates that the information present in one random variable about the other random variable is high. On the other hand, a small value of mutual information indicates that the random variables carry little information about the other. The concept of mutual information can be used to probe dependencies between two sets of data. A useful feature of mutual information is that there are no assumptions involved about the nature of the data.

Mutual Information compared with Correlation Coefficient

Correlation coefficient, which is another popular measure used to probe dependencies between two variables, measures only linear dependencies or second order statistics between

random variables. Mutual information is more general in application that it can probe non-linear statistical dependencies between random variables. Mutual information is invariant to component-wise monotonic transformations, that is $I(f(X), g(Y)) = I(X, Y)$ where $f(x)$ and $g(x)$ need not be linear functions, they can be any monotonic differentiable functions. On the other hand correlation coefficient is invariant only to component wise linear transformations.

Calculation of Mutual Information

Let us now consider the calculation of mutual information using a quantitative approach that is suitable for practical applications. The calculation of $I(X, Y)$ involves the calculation of individual probability distributions of X and Y as well as their joint probability distribution. $I(X, Y)$ can be computed as follows:

$$I(X, Y) = \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \cdot \log(p(x_i, y_j) / p(x_i)p(y_j)) \quad (2.17)$$

where $p(x_i)$ and $p(y_j)$ are probability measures for random variables X and Y respectively and $p(x_i, y_j)$ is the joint probability distribution for the pair of random variables (X, Y) .

Concept of Joint Mutual Information

Another interesting and very useful concept in information theory is the joint mutual information. Joint mutual information and mutual information are similar, except that joint mutual information is a more general concept and can be used to probe the dependencies

between a random vector and a random variable. Consider a random vector \bar{X} and a random variable Y . The joint mutual information between \bar{X} and Y is given by:

$$I(\bar{X}, Y) = \sum_i \sum_j p(\bar{x}_i, y_j) \cdot \log(p(\bar{x}_i, y_j) / p(\bar{x}_i) p(y_j)) \quad (2.18)$$

It is important to keep in mind that \bar{X} is a random vector and hence $p(\bar{x})$ corresponds to a joint distribution of all the random variables that constitute the random vector. Suppose \bar{X} is a two dimensional random vector given by $\bar{X} = (X_1, X_2)$, then $p(\bar{x})$ corresponds to the joint distribution $p(x_1, x_2)$. In this case this results in one more summation to appear in the calculation of joint mutual information. As is evident the calculation of joint mutual information becomes very much involved and computational heavy with the increase in the length of the random vector. But it is still practical for simulation purposes.

The joint mutual information is always greater than or equal to the maximum of the individual mutual informations involved. Consider the above random vector $\bar{X} = (X_1, X_2)$, then it follows that:

$$I(\bar{X}, Y) = I(X_1, X_2) = I(X_1, Y) + I(X_2, Y | X_1) = I(X_2, Y) + I(X_1, Y | X_2) \quad (2.19)$$

Clearly $I(\bar{X}, Y) \geq I(X_1, Y)$ and $I(\bar{X}, Y) \geq I(X_2, Y)$ since $I(X_2, Y | X_1)$ and $I(X_1, Y | X_2)$ are always nonnegative. Thus we can write $I(\bar{X}, Y) \geq \max(I(X_1, Y), I(X_2, Y))$.

Applications of Joint Mutual Information

The usefulness of joint mutual information becomes clear when it is necessary to probe the dependency of a target variable on a set of multiple feature variables. It must be noted that mutual information can only probe the dependency of a target variable on an individual feature variable.

To get a better insight into how joint mutual information works consider a target variable that depends on multiple feature variables. These multiple feature variables form a random vector. Obviously, to accurately specify the target variable we need all the feature variables. If we wish to quantify how individual feature variables affect the target variable, then we calculate mutual information between the target variable and each feature variable. Then the feature variable with maximum mutual information value is most important to predict the target variable. But if we have to quantify how a set of feature variables affects the target variable we compute the joint mutual information between the target variable and different vectors consisting of different sets of feature variables. It is worth noting that the joint mutual information thus calculated would always be greater than or equal to the maximum mutual information among the corresponding individual feature variables.

Suppose the task is to find the minimum number of feature variables required to predict the target variable within some acceptable error. Joint mutual information is a very useful tool for this kind of applications. The approach is to form a number of random vectors using different number of feature variables. Next we compute joint mutual information between each random vector and the target variable. If the joint mutual information statistics saturate, we pick the random vector for which joint mutual information almost saturates. The number of feature vectors present in the chosen random vector gives the minimum number

of feature vectors needed. On the other hand if the joint mutual information statistics do not saturate then it implies that all the feature vectors are essential to predict the target variable.

Bilmes [5] used mutual information between two variables on non-labeled data to reveal the mutual dependencies between the two components of the spectral energies in time and frequency and found that the information appears to be spread over relatively long temporal spans. Estimation of the joint distribution of feature vectors given a particular acoustic model is important in maximum-likelihood based speech recognition systems. In [6], it was shown that this task could be performed with better accuracy by modeling the joint distribution of time-localized feature vectors along with statistics relating those feature vectors to their surrounding context.

In [6], it was suggested that statistics corresponding to spectro-temporal loci in speech with relatively large mutual information are most useful in estimating the information contained in the feature-vector joint distribution and such statistics are most likely to generalize. These hypotheses were verified using both overlap plots and speech recognition word error results by using an EM algorithm to compute the mutual information between pairs of points in the time-frequency grid.

Morris [31] used mutual information to find the distribution of phonetic information across the on/off aligned auditory spectrogram of a speech database consisting of vowel-plosive-vowel (VPV) utterances. Automatic recognition was then used to test to what extent small high information samples are sufficient for plosive discrimination. The main idea in this work was that on and off events could be used to focus phoneme recognition on temporal intervals in the auditory nerve signal, which contain the greatest concentration of phonetic information. The information theoretic measure, Mutual information was used to

measure the distribution of the phonetic information present in a speech database of vowel-plosive-vowel utterances, which have been registered at on/off positions. A small number of samples corresponding to high information region were alone used to test speech recognition of the plosives.

The speech database consisted of 3000 vowel-plosive-vowel (VPV) signals sampled at 16 kHz. A spectrogram $S(f = 1\dots 15, t = 1\dots T)$ was obtained for each signal, by using a 256 sample short-term window, performing Hamming window weighting, followed by a discrete Fourier transform and critical band grouping into 15 bands using $Bark = 6.7 * asinh((Hz - 20) / 600)$ and logarithmic compression (Bark to dB). Each VPV signal was then normalized to have a mean energy of 0 dB and a standard deviation of 1.

On/Off positions were found for each signal and each variable sized VPV spectrogram matrix $S_i(f = 1\dots 15, t = 1\dots T_i)$ was projected onto a fixed size matrix $M_i(f = 1\dots 15, t = 1\dots 70)$ by superimposing all On/Off positions at fixed points in time ($t = 10, 30, 40, 60$) and linearly interpolating between these points. Each sample $X_i = M_i(f, t)$ was discretized over the range $\langle 1\dots R_x \rangle$, where $R_x = 8$. The corresponding phoneme index range $\langle 1\dots R_y \rangle$ was denoted by Y_i . Entropy estimates $H(X)$, $H(Y)$ and $H(X, Y)$ were computed. The mutual information denoted by $I_{f,t}(X, Y)$, was computed as:

$$I_{f,t}(X, Y) = [H(X) + H(Y) - H(X, Y)] / \log_2(\min(R_x, R_y)) \quad (2.20)$$

Information for initial and final vowels was found to be concentrated in time near to energy maxima and mostly in mid-range frequencies, whereas for plosives, it was

concentrated around consonantal closure and burst release positions and high frequency range. Also the joint mutual information for every 2-point sample in the spectrogram was computed. It was found that significant plosive information exists exclusively in joint patterns which link initial vowel and consonantal closure, closure with burst, closure and burst with a region in the center of the closure-burst interval and burst with final vowel.

Yang *et al.* [44] used mutual information to study the distribution in time and frequency of information relevant for phonetic classification. Short-term critical-band logarithmic energy $X(f_k, t)$ was used to represent information in time and frequency, where f_k denotes a particular frequency band and t denotes particular instant of time. Acoustic features $X(f_k, t)$ were derived from a short-time analysis of speech with a 20ms Hamming window advanced in 10ms steps, by computing squared magnitude *FFT* using a critical-band spaced weighting function. The speech database consisted of about 3 hours of telephone speech from the English portion of the Oregon Graduate Institute (*OGI*) multilingual database, with 50 seconds of extemporaneous speech from each of 210 different speakers. The speech database was hand-labeled using 19 commonly occurring phonemes. Thus, phonetic variable denoted by Y took on 19 values.

They computed mutual information between a point of logarithmic energy ($X(f_k, t)$) and phonetic variable (Y), denoted by $I(X(f_k, t), Y)$. A plot of $I(X(f_k, t), Y)$ for different values of f_k was plotted that showed how the information relevant for phonetic classification is distributed among the various bands of frequency. It was found that frequency bands carried information about the underlying phoneme, with dominant information in the frequency band ranging from 350-1300 Hz.

On the other hand a plot of $I(X(f_k, t-d), Y)$ for different delay values d , (where f_k is fixed) was plotted that gave insight into how the information is distributed in time. It was found that considerable information about a time frame is spread around 100 ms on either sides of the time frame, so contextual information in a window of about 100 ms to either side of the frame to be classified might be used. They also computed the joint mutual information between two points of logarithmic energy and phonetic variable, denoted by $I(X(f_i, t), X(f_k, t); Y)$. First, they fixed f_i and plotted $I(X(f_i, t), X(f_k, t); Y)$ for different values of f_k that gave them insight into how much information was gained by including another frequency band. On an average, they found that using an additional second frequency band resulted in around 0.25 bits increase in the information than in the case of using a single frequency band. In general the inclusion of a second point always provided information in addition to that provided by a single point.

They plotted $I(X(f, t), X(f, t-d); Y)$ (here both f_i and f_k were fixed and equal) for different delay values d showing how much information was gained by including another time frame or frames in addition to the current time frame. It was found that the spread of information in time was asymmetric in time with most of supporting information being found between 20 and 80 ms beyond the current frame.

The information theoretic approach discussed above is employed in the present work. Chapter 4 discusses the information theoretic approach in greater depth and presents various implementation details regarding the present work.

CHAPTER THREE

INTELLIGIBILITY OF SPEECH WITH HOLES IN THE SPECTRUM

3.1 Chapter Outline

This chapter explores the perception of speech with holes in the spectrum. Section 3.2 presents the motivation for current study and gives an insight into the concept of holes in the spectrum. Section 3.3 gives the various details about the experimental method used for conducting intelligibility tests. In Section 3.4 we present the results for both consonant and vowel stimuli. A detailed discussion of results is given in Section 3.5.

3.2 Motivation

Although many studies investigated the intelligibility of high-pass, low-pass [16] [24] [33] and band-pass filtered speech [39] [43], not many studies investigated the perception of band-stop filtered speech (i.e., speech with “holes” in the spectrum) or speech composed of disjoint frequency bands.

Lippmann [25] investigated the intelligibility of consonants that had a single “hole” in the middle of the spectrum. High consonant intelligibility (~90% correct) was maintained even after removing speech energy in the middle frequencies (800 to 4 kHz). Similar findings were reported by Riener *et al.* [35]. The intelligibility of sentences through single one-third octave bands centered around either 370 Hz or 6 kHz was roughly 23% when presented alone, but increased to 77% correct when presented simultaneously.

Both studies demonstrated that having access to low and high frequency information enabled listeners to identify speech with relatively high accuracy. Listeners seemed to “fill in” the missing speech information.

Recently, Shannon *et al.* [37] assessed the impact of the size and location of spectral holes with cochlear implant and normal hearing listeners. Results showed that holes in the low frequency region are more damaging than the holes in the high frequency region on speech recognition.

The aforementioned studies examined speech recognition either for a single hole varying in frequency location and size or for a single hole in the middle of the spectrum. The scope of the above studies was limited in the sense that it did not consider perception of speech composed of multiple disjoint bands involving low, middle and/or high frequency information. The present study addressed this question in a systematic fashion considering all possible combinations of missing disjoint bands from the spectrum.

The answer to the question of how listeners use and combine information across frequency regions, whether isolated or disparate, is not only important for understanding speech perception in general but it is also important for understanding speech perception by cochlear implant listeners.

Cochlear implants are based on the idea that there are surviving neurons in the vicinity of implanted electrodes. Cochlear implant listeners receive only a small number (4-6) of channels of frequency information, despite the fact that some implant processors transmit as many as 20 channels of information [14] [11]. There are currently three standing hypothesis on why implant listeners do not receive all the information.

The first hypothesis suggests that it is due to channel interaction which can potentially distort spectral information, the second hypothesis suggests that is because of warping in the spectral-tonotopic mapping [17], and the third hypothesis suggests that is because of the absence of neurons left for stimulation in the vicinity of the electrodes.

The lack of hair cells and/or surviving neurons in certain areas of the cochlea essentially creates holes in the spectrum. The extent of the effect of the holes in the spectrum on speech understanding is not well understood. It is not known, for instance, whether the spectral “holes” can account for some of the variability in performance among CI listeners.

It is therefore of interest to first find out which set of hole pattern(s) is most detrimental for speech recognition. The answer to that question would then be useful for determining ways to somehow make up for the lost information.

The aim of this study is to examine the effect of the location and size of spectral “holes” on vowel and consonant recognition. Understanding this effect will provide us with some insights as to why some CI listeners are not performing well, despite the wealth of information they receive.

In addition, we could use the data of this study to derive frequency importance function that takes into account the fact that listeners could combine information from disparate frequency regions in the spectrum. Complete details regarding this topic are given in the next chapter.

In the current study, speech is processed through six frequency bands, and synthesized as a sum of sine waves, with amplitudes equal to the root-mean-square (rms) energy of each frequency band, and frequencies set equal to the center frequencies of the

band-pass filters. Six channels were used as previous studies [27] reported that six channels were enough to achieve high levels of speech understanding.

To synthesize speech with a “hole” in a certain frequency band, we set the corresponding sine wave amplitude to zero. We systematically created “holes” in each of the six frequency bands (one “hole” at a time) and examined vowel and consonant recognition. Similarly, we synthesized speech with two “holes” in the spectrum, by setting the corresponding sine wave amplitudes to zero. All possible combinations were created including the scenarios where two “holes” were in adjacent frequency bands (thus making one large “hole”) or where the two “holes” were in disjoint frequency bands.

The intelligibility of speech having either a single “hole” in various bands or having two “holes” in disjoint or adjacent bands in the spectrum is assessed in this study with normal-hearing listeners. The extent of the effect of the location, size and pattern of spectral “holes” on vowel and consonant recognition is evaluated.

3.3 Method

3.3.1 Subjects

Twenty normal-hearing listeners participated in this study. All subjects were native speakers of American English. The subjects were paid for their participation. Eleven of the subjects were tested at University of Texas at Dallas and the remaining nine subjects were tested at Arizona State University. All the subjects participated in the study were from an age group of 20 to 25 years of age.

3.3.2 Speech Material

Subjects were tested on consonant and vowel recognition. Consonant data used for this experiment consisted of sixteen consonants in /aCa/ context taken from the Iowa consonant test [42]. The sixteen tokens of consonant data used for testing consisted of consonants in the syllables: “apa, ata, aka, afa, asa, asha, aba, ada, aga, ava, atha, aza, aja, ala, ama, ana.” All the syllables were produced by a male speaker.

Vowel data used consisted of vowels in the words: “heed, hid, hayed, head, had, hod, hud, hood, hoed, who’d, heard” produced by a male and a female speaker. A total of 22 vowel tokens were used for testing, 11 produced by 7 male speakers and 11 produced by 6 female speakers. It should be noted that not all speakers produced all 11 vowels. The stimuli were drawn from a set used by Hillenbrand *et al.* [21]. The first two formant frequencies (as estimated by Hillenbrand *et al.*) of the vowels used for testing are given in Table 3.1.

		had	hod	head	hayed	heard	hid	heed	hoed	hood	hud	who’d
F1	Male	627	786	555	438	466	384	331	500	424	629	319
	Female	666	883	693	492	518	486	428	538	494	809	435
F2	Male	1910	1341	1851	2196	1377	2039	2311	868	992	1146	938
	Female	2370	1682	1991	2437	1604	2332	2767	998	1102	1391	1384

Table 3.1: The first two formant frequencies (in Hz) of the male and female vowels used in this study.

3.3.3 Signal Processing

Speech material was first low-pass filtered using a sixth order elliptic filter with cut-off frequency 6000 Hz. Filtered speech was passed through a pre-emphasis filter with a cut-off

frequency of 2000 Hz. This was followed by band-pass filtering into six different frequency bands using sixth-order Butterworth filters with center frequencies of 394 Hz, 639 Hz, 1038 Hz, 1685 Hz, 2735 Hz and 4442 Hz respectively. The frequency boundaries of the six bands are given in Table 3.2.

Band	Lower Frequency (Hz)	Upper Frequency (Hz)	Center Frequency (Hz)
1	300	487	393
2	487	791	639
3	791	1284	1037
4	1284	2085	1685
5	2085	3388	2736
6	3388	5500	4444

Table 3.2: The 3-dB frequency boundaries of the 6 bands with the corresponding center frequencies (Hz) of each band.

The filters were designed to span the frequency range from 300 Hz to 5500 Hz in a logarithmic fashion. The output of each channel was passed through a rectifier followed by a second order Butterworth low-pass filter with a center frequency of 400 Hz to obtain the envelope of each channel output.

Corresponding to each channel a sinusoid was generated with frequency set to the center frequency of the channel and with amplitude set to the root-mean-squared (rms) energy of the channel envelope estimated every 4 ms. The phases of the sinusoids were estimated from the FFT of the speech segment. The sinusoids of each band were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment. A block diagram representing the signal processing tasks performed is given in Figure 3.1.

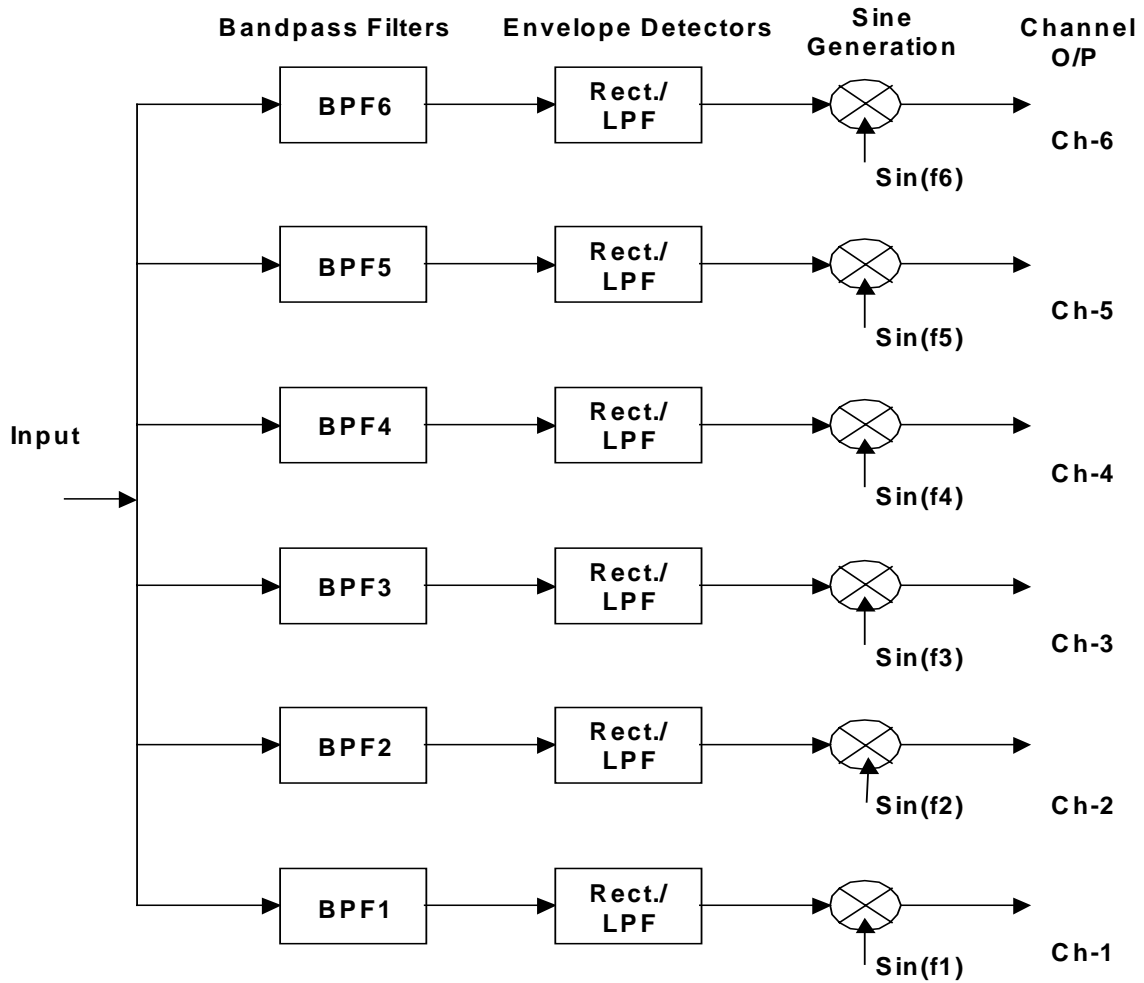


Figure 3.1: Block diagram representing signal processing performed.

To create a “hole” in frequency band N ($1 \leq N \leq 6$), we set the amplitude of the sinusoid corresponding to frequency band N to zero. Speech was synthesized using the remaining 5 channel amplitudes. Similarly, to create a “hole” in frequency band M and a

hole in frequency band N, we set the amplitudes of the sinusoids corresponding to frequency bands M and N to zero. Speech was synthesized using the remaining 4 channel amplitudes.

Vowel and consonant stimuli were created for 6 single “hole” conditions and 15 two “hole” conditions as shown in Table 3.3. All possible combinations of removing two out of the six frequency bands were considered. For comparative purposes, we also created a baseline condition in which we did not remove any frequency bands. Overall, subjects were tested with a total of 22 conditions.

3.3.4 Procedure

The experiments were performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones. The words were displayed on a computer monitor and a graphical user interface was used that enabled the subjects to indicate their response by clicking a button corresponding to the word played. No feedback was given during the test.

At the beginning of each test the subject was presented with a practice session in which the vowels or consonants were processed through 6 channels – no “holes” were introduced (baseline condition). After the practice session, the subjects were tested with the various spectral “hole” conditions.

Two groups of subjects were used, 11 from University of Texas-Dallas and 9 from Arizona State University. The eleven subjects at The University of Texas at Dallas were tested with the 14 test conditions labeled as 0, 1, 2, 3, 4, 5, 6, 7, 9, 15, 16, 18, 20 and 21 in Table 3.3. The 0-th condition was the baseline condition in which all 6 channels were

present. The nine subjects at Arizona State University were tested with the 9 conditions labeled as 0, 8, 10, 11, 12, 13, 14, 17 and 19 in Table 3.3.

Condition	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Channel 6
0	1	1	1	1	1	1
1	0	1	1	1	1	1
2	1	0	1	1	1	1
3	1	1	0	1	1	1
4	1	1	1	0	1	1
5	1	1	1	1	0	1
6	1	1	1	1	1	0
7	0	0	1	1	1	1
8	0	1	0	1	1	1
9	0	1	1	0	1	1
10	0	1	1	1	0	1
11	0	1	1	1	1	0
12	1	0	0	1	1	1
13	1	0	1	0	1	1
14	1	0	1	1	0	1
15	1	0	1	1	1	0
16	1	1	0	0	1	1
17	1	1	0	1	0	1
18	1	1	0	1	1	0
19	1	1	1	0	0	1
20	1	1	1	0	1	0
21	1	1	1	1	0	0

Table 3.3: The 22 test conditions considered in this study. The 0-th condition corresponds to the baseline condition. The channel(s) removed in each condition are indicated with a zero.

Note that both groups of subjects were tested with the baseline condition. The order in which the conditions were presented was counterbalanced between subjects to avoid order effects. In the vowel and consonant tests, there were 9 repetitions of each vowel and each consonant. The vowels and the consonants were completely randomized.

3.4 Results

3.4.1 Single –Hole Conditions

The mean percent correct scores for the single-“hole” conditions are shown in Figure 3.2.

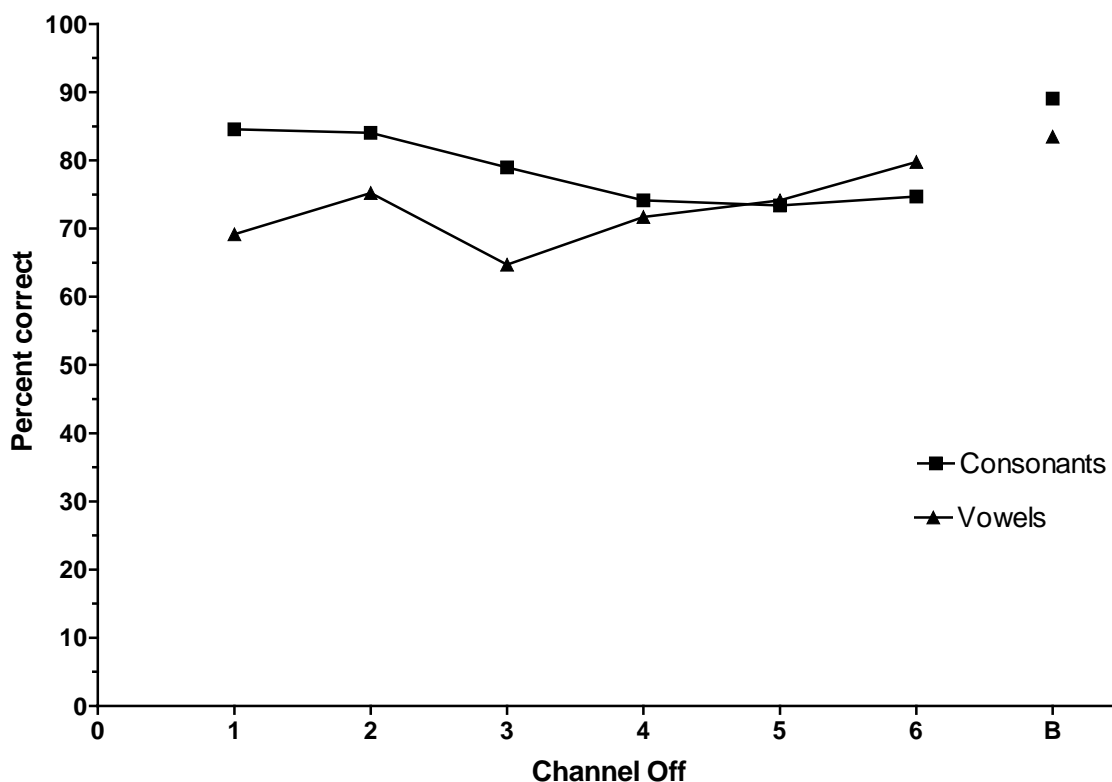


Figure 3.2: Mean percent scores for vowel and consonant recognition as a function of the location of the spectral “hole”. The “holes” were centered around the channel center frequencies. In the baseline condition, all channels were present.

A one-way ANOVA with repeated measures showed a significant main effect of the location of the spectral “hole” [$F(6,79) = 5.475, p < 0.05$] on consonant recognition. Post-hoc tests according to Tukey showed that the scores obtained with channels 4, 5 or 6 off were significantly lower from the baseline scores ($p < 0.05$). The average scores on the baseline condition were not significantly different ($p = 0.313$) between the two groups of subjects.

The scores obtained with channels 1, 2 or 3 off were not significantly different from the baseline scores. This outcome was not surprising since the major cues for consonant perception are known to be in the high frequencies.

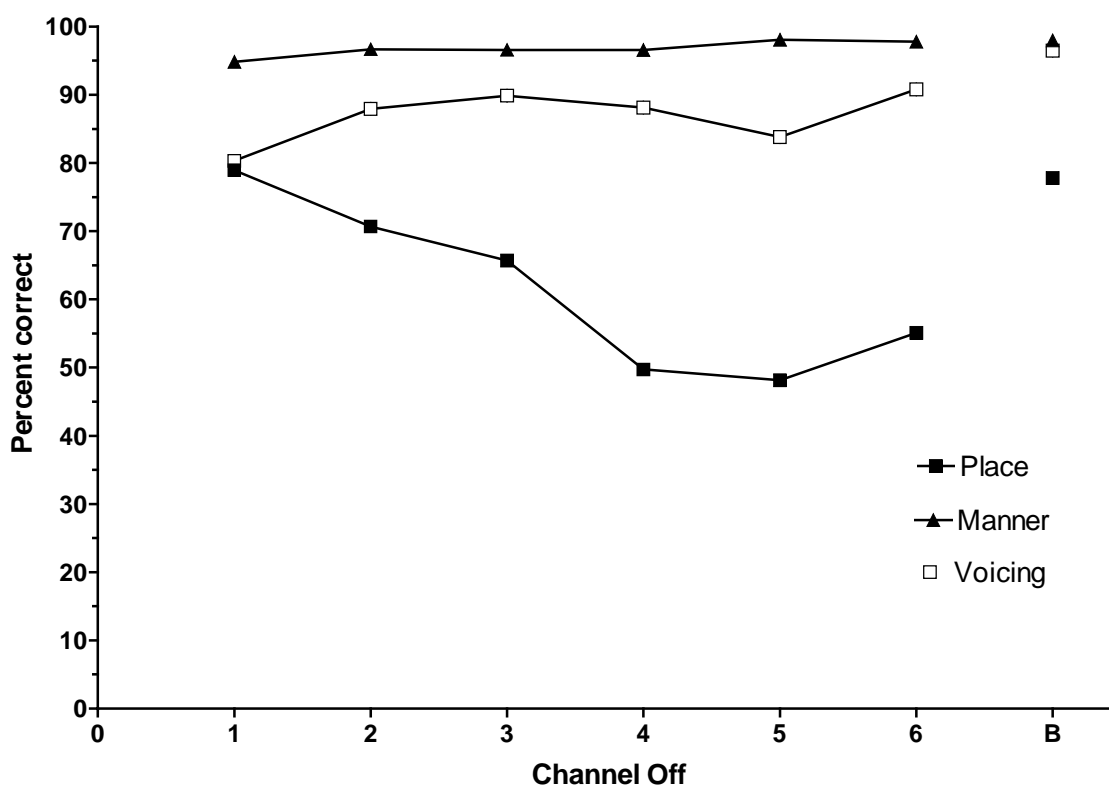


Figure 3.3: Percent information transmitted for the features place, manner and voicing as a function of the location of the spectral “hole”.

We further analyzed the consonant confusion matrices in terms of percent information transmitted as per Miller and Nicely [30]. The feature analysis is shown in Figure 3.3.

A one-way ANOVA showed a non-significant effect [$F(6, 70) = 0.664$ ($p=0.679$)] for the feature “manner” and a non-significant effect for the feature “voicing” [$F(6, 70) = 1.366$ ($p=0.241$)]. Voicing and manner did not seem to be affected by the location of the spectral “hole”.

We did find, however, that the feature “place” was significantly [$F(6, 70) = 11.088$ ($p<0.05$)] affected. A post-hoc Tukey test showed that conditions in which channel 4, 5 or 6 were off were significantly different from the baseline condition. This seems to be in agreement with the mean percent correct scores.

For the vowel data, a one-way ANOVA showed a significant main effect” [$F(6,79) = 5.724$, $p<0.05$] of the location of the spectral “hole” on vowel recognition. A post-hoc Tukey test showed that the scores obtained with channels 1, 3 or 4 off were significantly different from the baseline score ($p<0.05$).

The score obtained when channel 2 was off was not significantly different from the baseline score. The fact that channels 1, 3 and 4 were found to have a significant effect on vowel recognition was not surprising as those channels cover the F1-F2 frequency range.

3.4.2 Two-Hole Conditions

The mean scores for the two-“hole” conditions are shown in Figure 3.4. As expected, the mean scores dropped significantly when we introduced a second “hole” in the spectrum.

The baseline score for consonant recognition, for instance, dropped from 89.06% to an average (across all conditions) of 69.6%.

A one-way ANOVA showed a significant main effect on consonant recognition when two “holes” were introduced in the spectrum [$F(15,168) = 6.904, p < 0.05$]. Post-hoc Tukey tests showed that several channel pair combinations were significantly affected consonant recognition and those pairs were: (1,2), (1,4), (1,6), (2,3), (2,6), (3,4), (3,6), (4,5), (4,6) and (5,6).

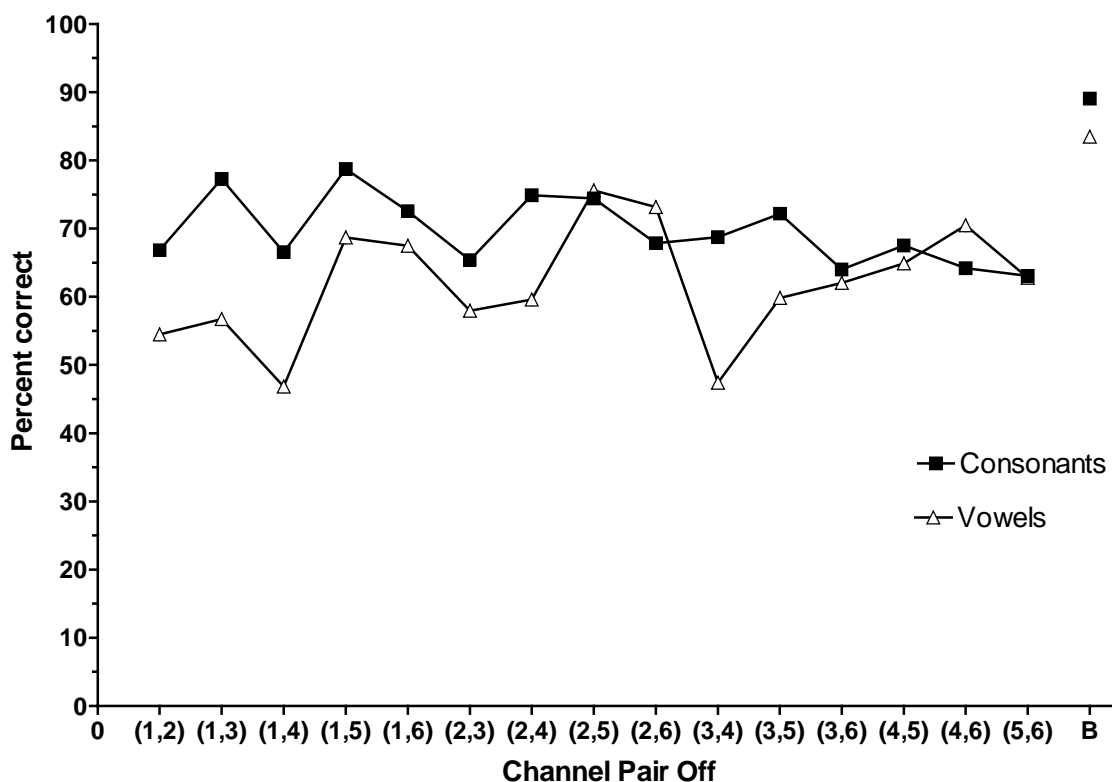


Figure 3.4: Mean percent scores for vowel and consonant recognition as a function of the location of the pair of spectral “holes”. The “holes” were introduced at frequencies centered at the channel pairs indicated. In condition (1,4), for instance, channels 1 and 4 were removed from the spectrum. In the baseline condition, all channels were present.

Overall, we found that the scores obtained with channel pairs that included channels 4, 5 or 6 were significantly lower from the baseline score ($p < 0.05$). This seems to be consistent with the single-“hole” conditions, and reinforces the message that channels 4, 5 and 6 are very important for consonant recognition.

The consonant confusion matrices were analyzed in terms of percent information transmitted. The feature analysis is shown in the figure 3.5.

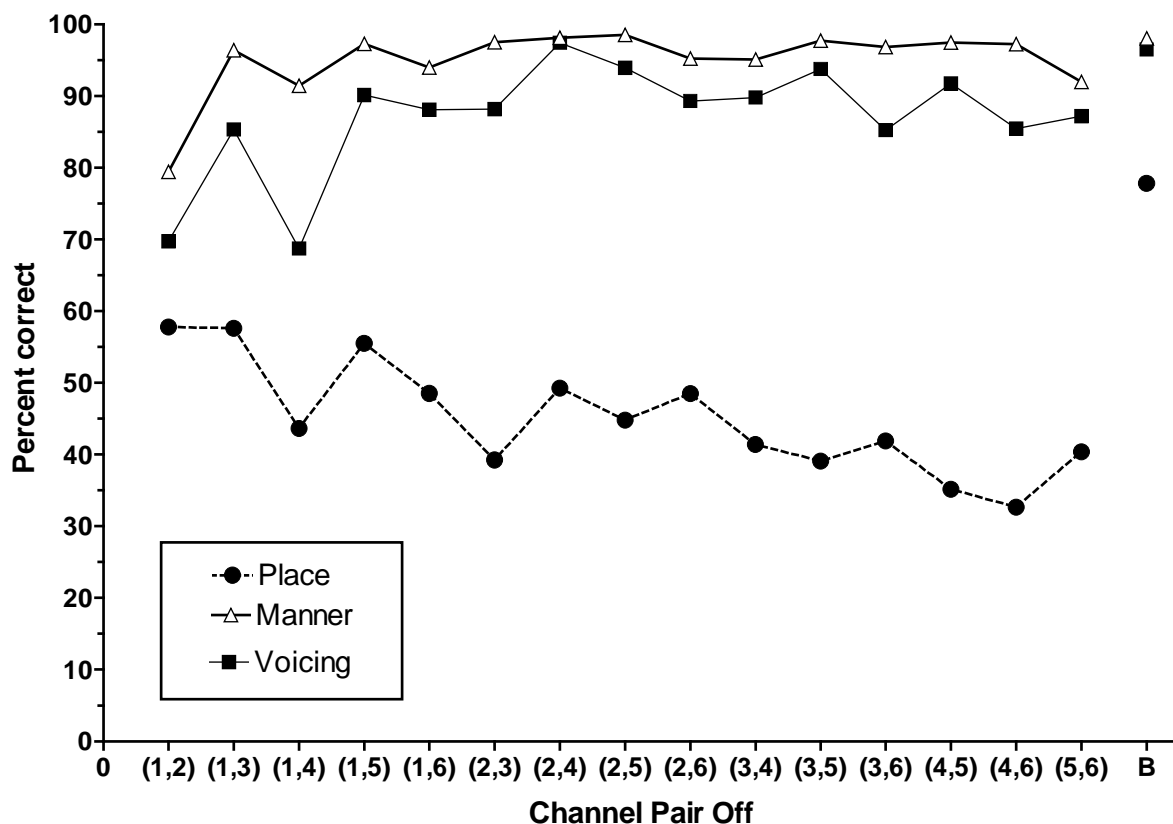


Figure 3.5: Percent information transmitted for the features place, manner and voicing as a function of the location of the pair of frequency bands removed.

A one-way ANOVA with repeated measures showed a significant effect [$F(15,120) = 5.5, p < 0.005$] for the feature “manner”, a significant effect for the feature “voicing”

[F(15,120) 3.5, $p < 0.0005$], and a significant effect [F(15,120) = 6.7, $p < 0.0005$] for the feature “place”.

Post-hoc Tukey tests showed that the “manner” score obtained with the channel pair (1,2) removed was significantly lower ($p < 0.0005$) than the baseline score. The “voicing” scores obtained with the channel pairs (1,2) and (1,4) removed were significantly lower ($p < 0.005$) than the baseline score. All “place” scores were significantly lower ($p < 0.005$) than the baseline score.

For vowel recognition, a one-way ANOVA showed a significant main effect [F(15,154) = 8.648, $p < 0.05$] when two “holes” in the spectrum were introduced. Post-hoc Tukey tests showed that several channel pair combinations were significantly affected on vowel recognition, and those pairs were: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4), (3,5) and (5,6). The drop in vowel performance when both channels 5 and 6 were removed was due to the low scores obtained for the vowels in “heed”, “hid” and “hayed”.

Post-hoc Tukey tests showed that the scores obtained with channel pairs that included channel 1, 3 or 4 were significantly lower from the baseline scores ($p < 0.05$), consistent with the outcome in the single-hole conditions. More specifically, the lowest scores on vowel recognition were obtained with channel pairs (1,2), (1,3), (1,4) and (2,4).

3.5 Discussion

The above results suggest that vowel and consonant recognition suffers when “holes” are introduced in the spectrum. The degree of degradation in recognition performance as well as effect of the location of the spectral “holes” was different for vowels and consonants.

3.5.1. Effect of location of spectral “holes”

For vowels, statistical analysis showed a significant drop in performance when either channels 1, 3 or 4, centered at 393 Hz, 1037 Hz, and 1685 Hz respectively, were removed. It is safe to assume that channel 1 codes F1 information, and channels 3 and 4 code F2 information for most vowels (Table 3.1). Channel 3 may also code F1 information for some female vowels (i.e., vowels in “hod” and “hud”).

Depending on how high the F2 frequency is for some speakers, channel 5 (and, indirectly, channel 6) may also be important for the recognition of some vowels. Channel 5 may, for instance, code F2 information for some vowels (i.e., “heed”, “hid”, “hayed”) produced by female speakers or children who generally have a high F2 frequency. Indeed, close examination of the individual vowel’s scores indicated that the identification of the female vowels in “heed”, “hid” and “hayed” dropped significantly when both channels 5 and 6 were removed.

It is interesting to note that vowel recognition performance was not significantly affected when channel 2 (centered at 639 Hz) was removed. Channel 2 most likely codes F1 information either together with channel 1 or alone. Information about F1 is captured by channel 2 alone when the first formant frequency of the vowel falls near the center frequency of channel 2. In that case, a peak in the channel spectrum is observed at channel 2, and consequently removing channel 2 will significantly reduce performance.

This is demonstrated in Figure 3.6, which shows the listeners’ individual vowel performance when channel 2 was removed. Vowels / ϵ / and / a / were the only vowels that were significantly affected because the F1 frequency of those vowels happened to be near the center frequency of channel 2. For the remaining vowels, however, as it is evident from

Figure 3.5, listeners seemed to infer F1 information from channel 1 when they did not have access to channel 2 information. Mean percent correct scores on individual consonant and vowel recognition for various conditions are presented in Appendix B.

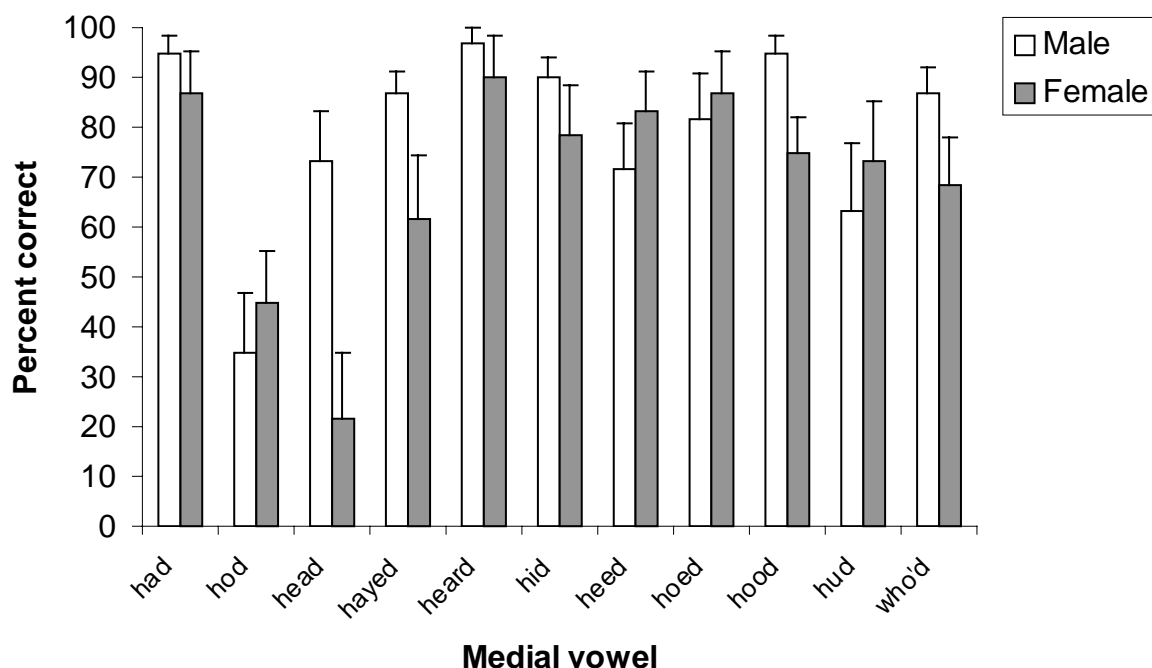


Figure 3.6: Mean percent scores on individual vowel recognition for the condition in which channel 2 was removed from the spectrum (n=20). The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively. Error bars indicate standard errors of the mean.

This suggests that having a rough estimate of F1 is sufficient for the recognition of most vowels. That was not the case with F2, since removing either channels 3 or 4 affected vowel recognition.

For consonants, statistical analysis showed a significant drop in performance when either channels 4, 5 or 6 were removed. This outcome is consistent with the conventional view that high frequency cues are important for recognition of “place”. So, the drop in

consonant recognition performance was primarily due to a reduction in information transmitted for “place” (Figure 3.3).

Removing any of the low frequency channels (1-4) affected vowel recognition, and removing any of the high-frequency channels (4-6) affected consonant recognition. Interestingly enough, channel 4, which had a center frequency of 1685 Hz, was found to be important for both vowel and consonant recognition.

The frequency (1685 Hz) corresponding to channel 4 is close to the well-known crossover frequency estimated in articulation index studies. Depending on the speech material used, the crossover frequency was found in articulation index studies to be in the range of 1550 Hz to 1900 Hz [16] [40].

Overall, with the exception of channel 2 which did not significantly affect either vowel or consonant recognition, removing single channels caused a modest, but significant, reduction in performance in vowel and consonant recognition. Consonant recognition was less affected than vowel recognition.

It should be noted that the drop in performance, although statistically significant, was not dramatic for either consonant or vowel recognition. Even in the worst-case conditions, vowel and consonant recognition remained about 70% correct. So, relatively high vowel and consonant recognition performance can be maintained even with a single “hole” in the spectrum.

This outcome is consistent with the data reported by Shannon *et al.* [37] with cochlear implant listeners. Shannon *et al.* artificially created single “holes” by turning off a number of (apical, middle or basal) electrodes in CI listeners who were fitted with the 22-electrode Nucleus device. Holes were created that were 2-8 electrodes wide corresponding

to 1.5-6.0 mm width. High vowel and consonant recognition was maintained even when as many as 4 electrodes were turned off either in the low, middle or high frequency regions.

3.5.2 Effect of size and pattern of spectral “holes”

In five out of the 15 conditions tested, the size of the “hole” or equivalently the width of the notch in the spectrum doubled, since in these conditions [i.e., channel pairs (1,2), (2,3), (3,4), (4,5) and (5,6)] the channels that were removed were adjacent to each other. This caused a large drop in vowel recognition performance, and only a moderate drop in performance for consonant recognition.

Vowel recognition dropped in some cases to as low as 47% correct. The lowest performance occurred when F1 information was missing (e.g., pair (1,2)), when F2 information was missing [pairs (2,3), (3,4)] or when both F1 and F2 information was missing [pair (1,4)].

Consonant recognition was only mildly affected by the location of the pairs of frequency bands removed. The decrease in consonant identification was due primarily to the loss of “place” information (Figure 3.5). The manner and voicing features were significantly affected only when information about F1 was missing.

Overall, consonant identification remained robust and hovered around 70% for most conditions. Even when the middle and high frequency bands were absent, consonant recognition remained around 70% correct.

This outcome is consistent with the data reported by Lippmann [25], who evaluated consonant recognition by presenting a low-pass band below 800 Hz and a high-pass frequency band with cutoff frequency varying from 3.15 kHz to 8 kHz. He observed a high

score of 91% correct when the high-pass cut-off frequency was 3.15 kHz. This corresponded to the case where channels 4 and 5 were removed in our study. The score for that condition was 67.5% correct. The difference in scores between our study and Lippmann's can be attributed to the fact that our listeners had only access to four channels (2 channels were removed) of frequency information.

Similar findings were reported by Dorman *et al.* [10] with CI listeners fitted with a 4-channel processor. No significant difference was found between the consonant identification score obtained with only channels 1 (low frequency) and 4 (high frequency) activated and the score obtained with all 4 channels activated.

Our study extended Dorman's and Lippmann's findings to show that high consonant recognition can be maintained even in the absence of not only middle frequencies but also low-high, low-middle, low-high and middle-high frequency information.

Overall, we can say that vowel recognition seems to be sensitive to the size and pattern of "holes" in the spectrum. This was not surprising, since it is known that listeners rely primarily on spectral cues to identify vowels. In contrast, listeners make use of both temporal-envelope cues and spectral cues to identify consonants. In the absence of sufficient spectral cues, listeners probably rely more on temporal cues to identify consonants.

As shown in this study (Figure 3.5), these temporal cues did not seem to be affected by the frequency location of the pair of bands removed [except when channels (1,2) were removed]. We believe that is the reason that consonant recognition remained relatively high (~70% correct) even when two "holes" were introduced in the spectrum. The above results have certain implications for cochlear implants.

The finding that the location and pattern of “holes” affects mostly vowel recognition suggests that in cochlear implants, neuron survival (responsible for the “holes” in the spectrum) ought to account for some of the variability in vowel recognition performance among CI listeners.

CHAPTER FOUR

FREQUENCY IMPORTANCE FUNCTIONS

4.1 Chapter Outline

Here we present a detailed discussion of the derivation of the frequency importance functions. First we present a least square approach that utilizes the data from the intelligibility tests discussed in previous section to assess the importance of various frequency bands to speech perception. A detailed discussion is given in section 4.2. Next we present an information theoretic analysis to arrive at the importance of the various frequency bands based on the computation of the mutual information between the spectral energies and the corresponding phonetic labels. Section 4.3 presents a detailed discussion regarding this topic.

4.2 Least Squares Approach

4.2.1 Limitations of Articulation Index (AI)

Several investigators have used the AI method to determine frequency importance functions. The AI method uses a quantity between zero and one to represent the proportion of speech information available in a specific frequency band to the listener. This information is then multiplied by a frequency-importance or “weighting” function, which is obtained using a rather time-consuming process of low-pass and high-pass filtering speech.

The AI method assumes that the information contained in each band is independent of the information contained in other bands and does not take into account the fact that listeners may combine speech information from multiple disjoint bands. This was first demonstrated by Kryter [24] who evaluated recognition of pass-band speech, and showed that the AI could not adequately predict intelligibility of pass-band speech. Similar findings were also reported by Grant and Braida [19].

Several methods were proposed in the literature to circumvent this shortcoming including the correlational method by Doherty and Turner [9] and a recent method based on statistical decision theory by Musch and Buus [32]. In this study, we use the data from intelligibility tests to derive a frequency importance function based on a least squares approach. Unlike the AI method, the proposed least squares approach makes use of the listeners' scores on perception of vowels and consonants composed of disjoint frequency bands.

4.2.2 Estimation of Weights

Our approach to obtain the importance of each frequency band follows the method proposed by Ahumada and Lovell [1]. We used the results from intelligibility tests for holes in the spectrum to predict the importance or perceptual “weight” of each channel.

We calculated the weight w_i of each channel by predicting the responses of the subject as a linear combination of the strength of each channel, i.e.,

$$R_k = \sum_{i=1}^6 w_i E_{ik} \quad (4.1)$$

where R_k is the mean percent correct score for condition k and E_{ik} is the strength of the i -th channel corresponding to condition k . The strength of each channel is a binary value that can be either 0 or 1 depending on whether the channel is off or on respectively. The value of k ranges from 1 to 22 spanning all channel combinations (Table 3.3). Forming the prediction error e_k as:

$$e_k = R_k - \sum_{i=1}^6 w_i E_{ik} \quad (4.2)$$

We can estimate the channel weights by minimizing the sum of all the squared errors with respect to w_i . Alternatively, Equation (4.1) can be written in matrix form as:

$$R = E \bullet W \quad (4.3)$$

where R is a 22-dimensional vector containing the mean percent correct scores for conditions 1 to 22, E is the data matrix (22x6) consisting of the strengths of each channel (Table 3.3) and $W=[w_1, w_2, \dots, w_6]$ is a 6-dimensional vector consisting of the desired channel weights.

The above set of equations represents an over deterministic system of equations since we have six unknowns (the channel weights) and 22 equations (one for each condition). We calculated the weights W by solving the matrix equation given by (4.3) using a least squares approach:

$$W = (E^T E)^{-1} E^T R \quad (4.4)$$

After obtaining the solution from Equation (4.4), we normalized the weights so that the sum of all the weights was equal to one.

4.2.3 Results

The relative weights of the various channels are shown in Figures 4.1 and 4.2 for the vowel and consonant stimuli.

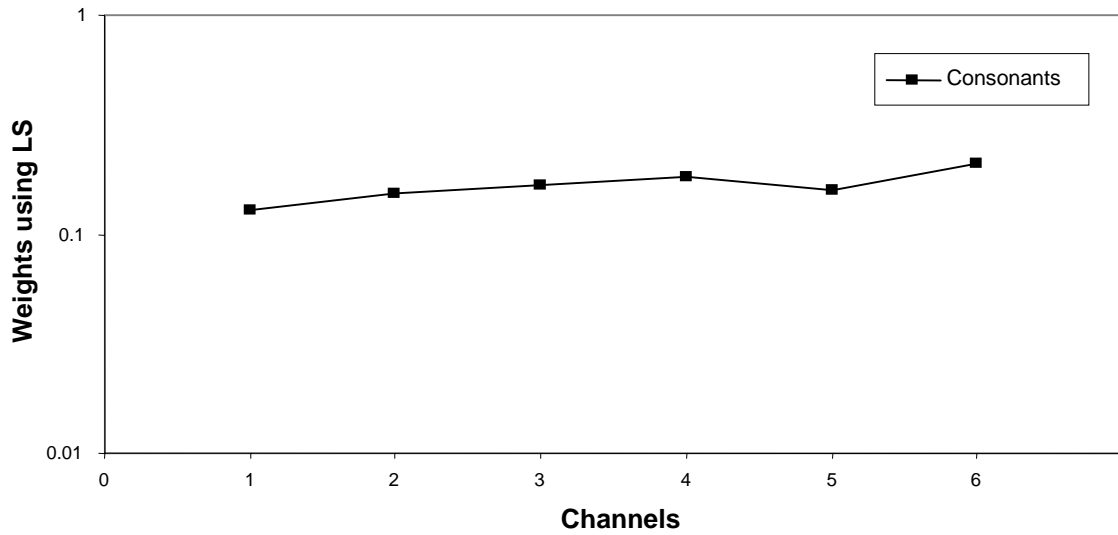


Figure 4.1: Frequency-importance function for Consonants.

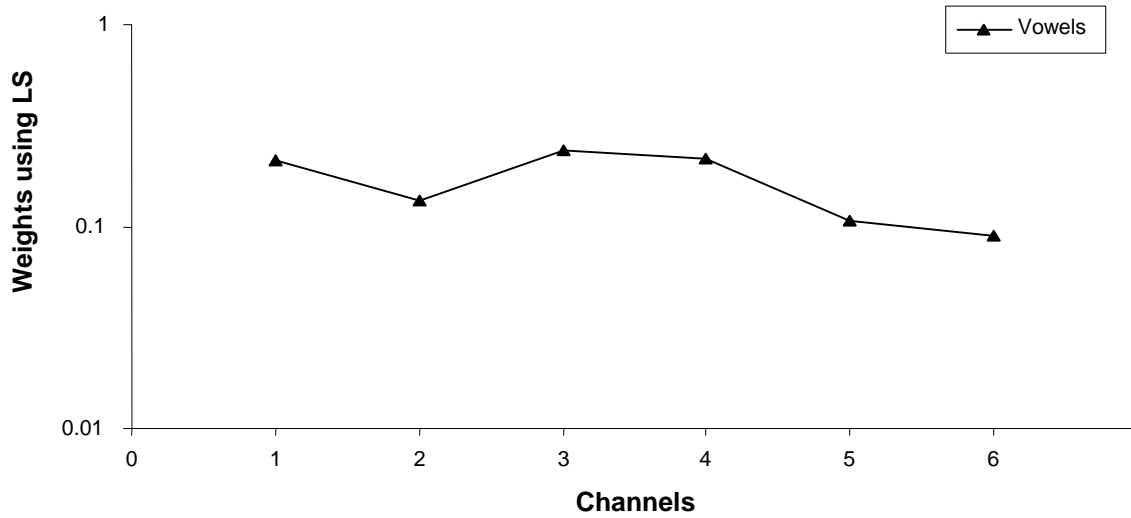


Figure 4.2: Frequency-importance function for Vowels.

As can be seen, the shape of the weighting function was different for vowels and consonants. For vowels, there was unequal weighting across the various channels suggesting that each channel contributed differently in understanding these vowel tokens. Channels 1, 3 and 4, centered at 393, 1037 and 1685 Hz respectively, received the largest weight. This outcome was consistent with the listener's reduction in performance in intelligibility tests when those channels were removed. Also, consistent with our data from intelligibility tests, channels 2, 5 and 6 received the lowest weight.

4.2.4 Discussion

The weighting function for the consonants was relatively flat. This suggests that for consonant recognition all channels are equally important. This outcome is consistent with the data reported recently by Mehr *et al.* [29]. Mehr *et al.* estimated the frequency importance function of nonsense syllables using the correlational method. Speech was divided into six frequency bands and a randomly chosen level of filtered noise was added to each channel on each trial. Channels in which the signal-to-noise ratio was more highly correlated with performance had a larger weight, and channels with smaller correlations had lower weights. Their results showed a flat weighting function for normal-hearing listeners. Unequal weighting functions, accompanied with a large variability among subjects, was noted for the CI listeners in their study.

The individual listener's weighting functions are given in Figure 4.3 for vowel and consonant recognition. Weighting functions are given for 6 of the 20 subjects, two subjects with the highest vowel scores (Figure 4.3, panels *a* and *b*), two with the middle vowel scores (panels *c* and *d*) and two with the lowest vowel scores (panels *e* and *f*).

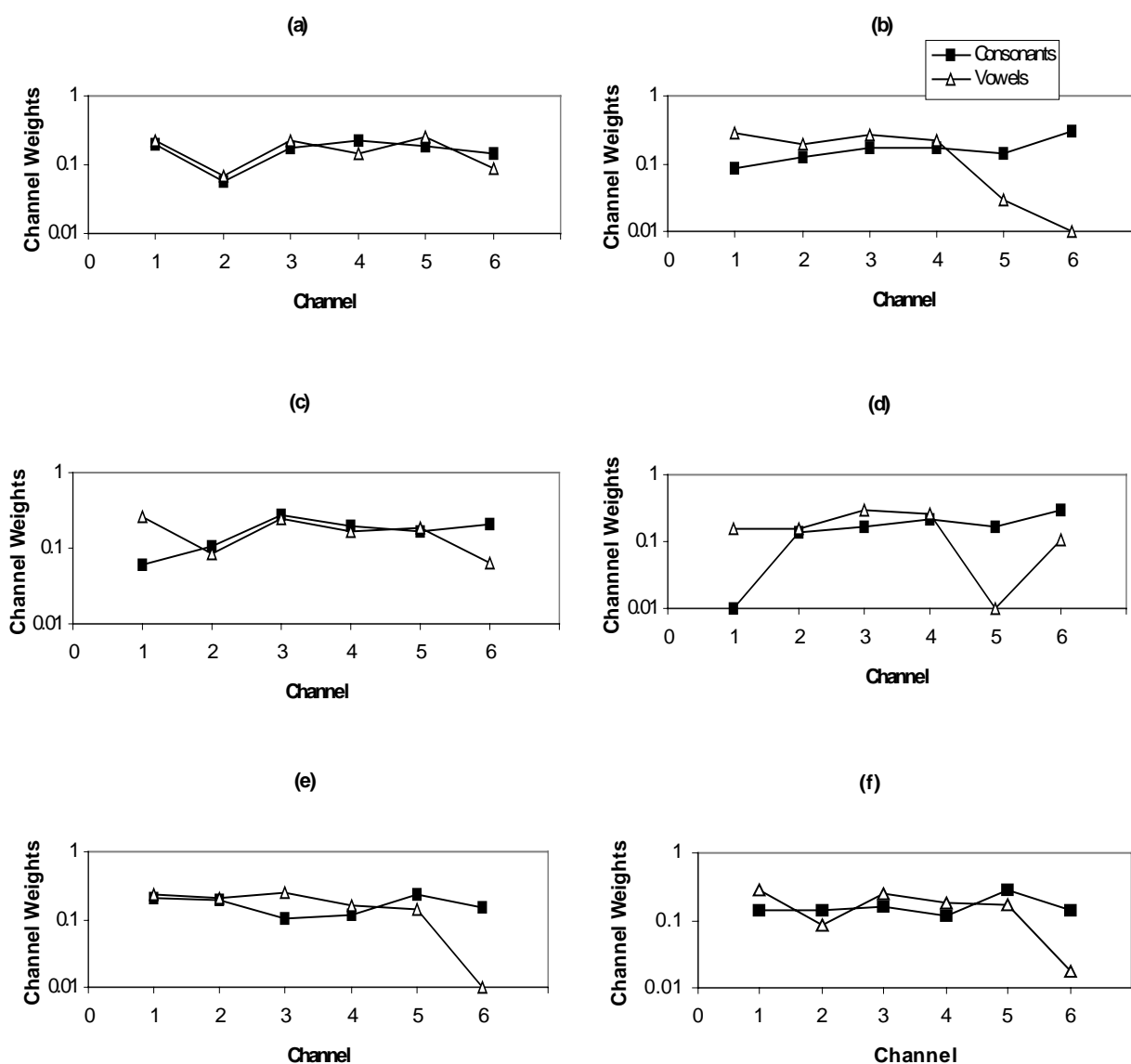


Figure 4.3: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for two subjects with the highest vowel scores, panels (c) and (d) show the functions for two subjects with middle scores, and panels (e) and (f) show the functions for two subjects with the lowest vowel scores.

Most listeners had a relatively flat weighting function for consonants with a small variability. There was a larger variability among subjects in the shape of the weighting

functions for vowels, suggesting that subjects used different listening strategies for vowel recognition. Individual listener's frequency-importance functions for vowel and consonant recognition for all the subjects are presented in Appendix A.

The fact that the weighting functions for vowels and consonants were different suggests that subjects were using different listening strategies to identify vowels and consonants. For vowel identification, listeners rely primarily on spectral cues and therefore place more emphasis or more "weight" on the channels that code F1 and F2 information. For consonant identification, listeners rely on both temporal-envelope and spectral cues, which are distributed across all channels. Hence, all frequency bands contributed equally to consonant identification, at least for the filter spacing used in this study. The data from Intelligibility experiments for holes in the spectrum are consistent with this conclusion.

The fact that consonant recognition remained relatively constant, around 70% correct, regardless of which pairs of channels were removed, clearly demonstrated that all channels contributed equally to consonant recognition. Had the listeners placed more emphasis on certain channels or pairs of channels, we would have seen a dramatic decrease in performance at those channel(s), as we did with the vowels.

We suspect that, in general, the frequency importance function must be dependent, among other factors, on the speech material and the frequency spacing used. Studebaker *et al.* [40], for instance, showed that the shape of articulation index function and the crossover frequency depended on the speech material.

We did not vary the frequency spacing in this study, but rather used the logarithmic spacing typically implemented in current cochlear implant processors. According to the data

obtained in this Experiment, logarithmic spacing provided equal amount of speech information in each frequency band for consonant identification.

This outcome has important implications for cochlear implants. Logarithmic spacing would be desirable assuming that CI listeners are able to extract information from *all* their electrodes. As shown by many investigators [14] [11] [45], that is not the case. This suggests that the frequency spacing should be customized for each CI subject in such a way that their resulting frequency-importance function has larger weights on the functional electrodes and smaller weights on the not-so functional electrodes.

Despite the differences between the least squares approach used in this study and the correlational method used by Mehr *et al.* [29], we obtained a similar (almost identical) weighting function for nonsense syllables. The testing process involved in deriving the weighting functions is time-consuming and therefore both methods are impractical for clinical applications. Another drawback of the correlational method is that it is dependent on the number of trials used for testing. As many as 1200 trials were required in some cases to get significant raw correlations [29] [41].

Our method is not largely dependent on the number of trials, but requires an adequate number of conditions. In our study, we needed to run a total of 22 conditions, which are considerably less than the 135 conditions needed for articulation index studies [40] to estimate the frequency importance function. In brief, the least squares approach proposed in this study is another viable approach for obtaining frequency-importance functions.

4.3 Information Theoretic Analysis

In our work, we seek to find how the information about speech recognition pertaining to the various phonemes is distributed in time and frequency. For this task we use the mutual information between phonetic labels and spectral energy. The analysis is performed in the perspective of cochlear implants in the sense how information is distributed among the various channels of the cochlear implant. The speech data is filtered into six frequency bands corresponding to six channels of the cochlear implant.

4.3.1 Processing of Speech Data

The speech data consisted of a sequence of speech files. Corresponding to each speech file we have a phonetic file. We process each speech file into six frequency bands as discussed in chapter 3. Thus we filter the input speech file into six channels using band-pass filters. This provides the basis for analyzing how information is distributed across different frequency bands. The corresponding phonetic file is read in parallel to the processing of speech file. This is essential since we are concerned with the calculation of mutual information between the spectral energy in each channel and phonetic labels.

4.3.2 Computation of Spectral Energy

In each channel, the spectral energy is computed for every 4ms frame. The means we use to calculate the spectral energy is the root mean square energy of each channel over 4 ms. For every 4ms frame the corresponding phonetic label is read from the phonetic file. Thus for every 4ms we compute six energy values corresponding to the six channels and read corresponding phonetic label. The spectral energies thus computed are stored sequentially in

a file. The corresponding phonetic label is also stored in another file. This procedure is repeated till the end of the speech file. All the speech files are processed in the same manner.

To calculate the mutual information between two variables we need to calculate the individual probability distributions of each variable as well as their joint distribution. To calculate the probability distribution of spectral energy we need to quantize the raw spectral energy computed as described above into a fixed number of bins. It must be noted that quantization process is inherent in practical implementations of many signal-processing systems.

4.3.3 Quantization of Spectral Energy

Here we quantize the spectral energy using uniform quantization. As a precursor to quantization we compute the 99.5 percentile of spectral energy to better utilize the dynamic range. We set this 99.5 percentile as the maximum for the quantization so that quantization is performed in a uniform way so that there is not a bias in the computation of levels of quantization. Next we proceed with quantization of the dynamic range of spectral energy. We use a 64 level uniform quantizer for this purpose. Thus the spectral energies corresponding to each channel are divided into 64 bins. We store the quantized spectral energies of each channel in terms of their bin numbers in a file to facilitate easy analysis.

4.3.4 Calculation of Probability Distributions

Next we proceed to calculate the probability distributions for the spectral energies of each channel. Probability distribution is calculated in terms of the bins of the quantizer. First we compute the number of frames of speech that fall into each bin. Then the probability of a

particular energy bin is obtained by dividing the number of speech frames that fall into that bin by the total number of speech frames. This is repeated for all the bins to obtain the complete distribution of spectral energy.

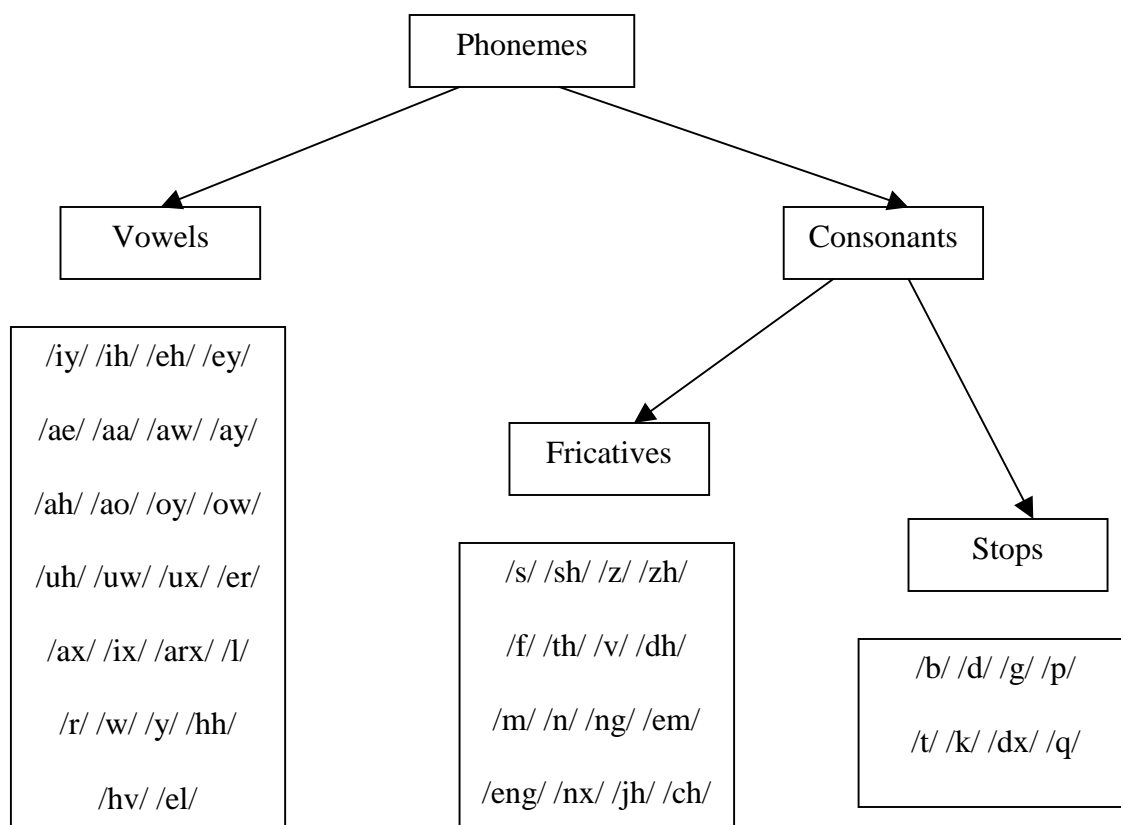


Table 4.1: Phonemes in American English.

We also calculate the probability distribution for the phonetic labels. We used 60 phonemes listed in the *TIMIT* database as the sample space for the phonetic labels. The complete list of phonetic labels is given in Table 4.1. The phonemes in American English are broadly classified into consonants and vowels. Consonants can be further sub-classified into stops and fricatives. First we compute the number of frames of speech that fall into each phonetic label. Then the probability of a particular phonetic label is obtained by dividing the

number of speech frames that fall in that phonetic label by the total number of speech frames.

4.3.5 Computation of Joint Distributions

The computation of joint distribution of spectral energies and phonetic labels is more involved. Here the sample space consists of the 64 quantized energy bins and the 60 phonetic labels. Thus the sample space is 64×60 joint space. More specifically a sample point corresponds to the occurrence of a particular energy bin coupled with the occurrence of a particular phoneme. We compute the probability of a sample point by counting the number of times a speech frame falls into a particular energy bin when a particular phoneme occurs and dividing by the total number of speech frames multiplied by total number of phonetic labels employed. This is computed for all the sample points to obtain the complete joint distribution between spectral energy and phonetic labels.

4.3.6 Calculation of Mutual Information

We compute the mutual information between spectral energies and phonetic labels by using the distributions of spectral energies, distribution of phonetic labels and joint distribution of spectral energies and phonetic labels according to formula given by Equation (2.17).

The calculation of mutual information is performed for each channel. Thus for six channels, we obtain six values of mutual information relating spectral energy in that channel to the phonetic labels that occurred in the speech database. We plot the values of mutual information across the six channels to observe how information is distributed among the various channels of the implant. In simple terms, a channel with a large mutual information

value is more important for speech perception than a channel with a small value of mutual information.

4.3.7 A Summary of Implementation Procedure

Finally we summarize the various implementation details involved in the calculation of mutual information and joint mutual information in a four-step process. We illustrate each step by using a sample speech data as input. The sample speech data consists of a single speech file ‘apa.wav’ and the corresponding phonetic file is ‘apa.phn’.

- Step1: Process the input speech data into six frequency bands and compute spectral energy for every 4 ms frame in each channel. Here input data corresponds to 162 frames.
- Step2: Quantize the frame energy into 64 levels. The plots of frame energy and quantized frame energy in each channel are shown below in Figure 4.4. The figure indicates that 64 level quantization preserves energy distribution.
- Step3: Compute the probability distribution for quantized energy bins corresponding to various frequency bands. A plot of quantized energy distribution for each channel is given in Figures 4.5 and 4.6. Also compute the distribution for the phonetic labels and the joint distribution between the spectral energies and the phonetic labels.
- Step4: Calculate mutual information using the various probability distributions. A plot of mutual information for the six channels employed in this study is given in Figure 4.7.

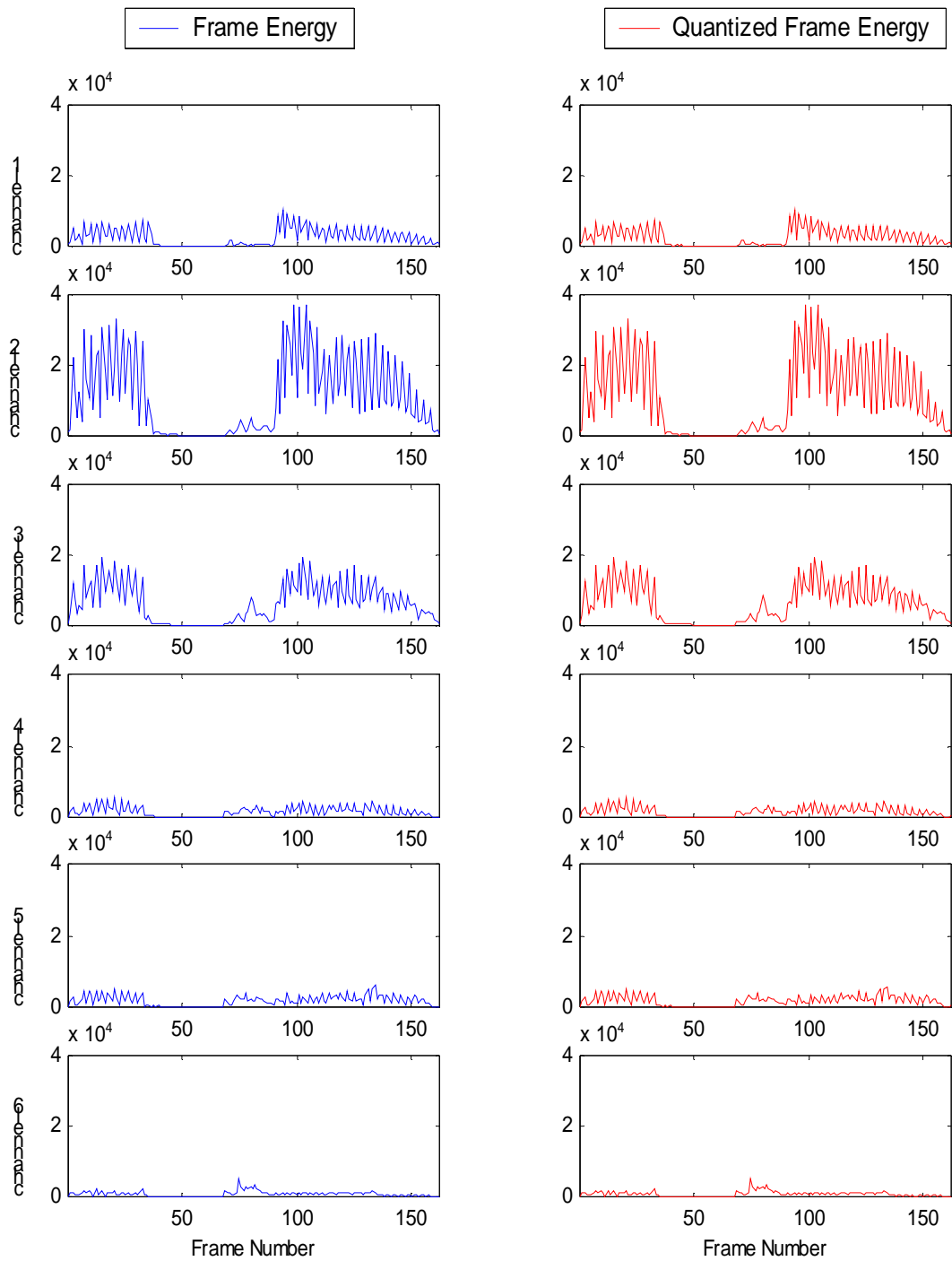


Figure 4.4: Quantized frame energy plots versus frame energy plots

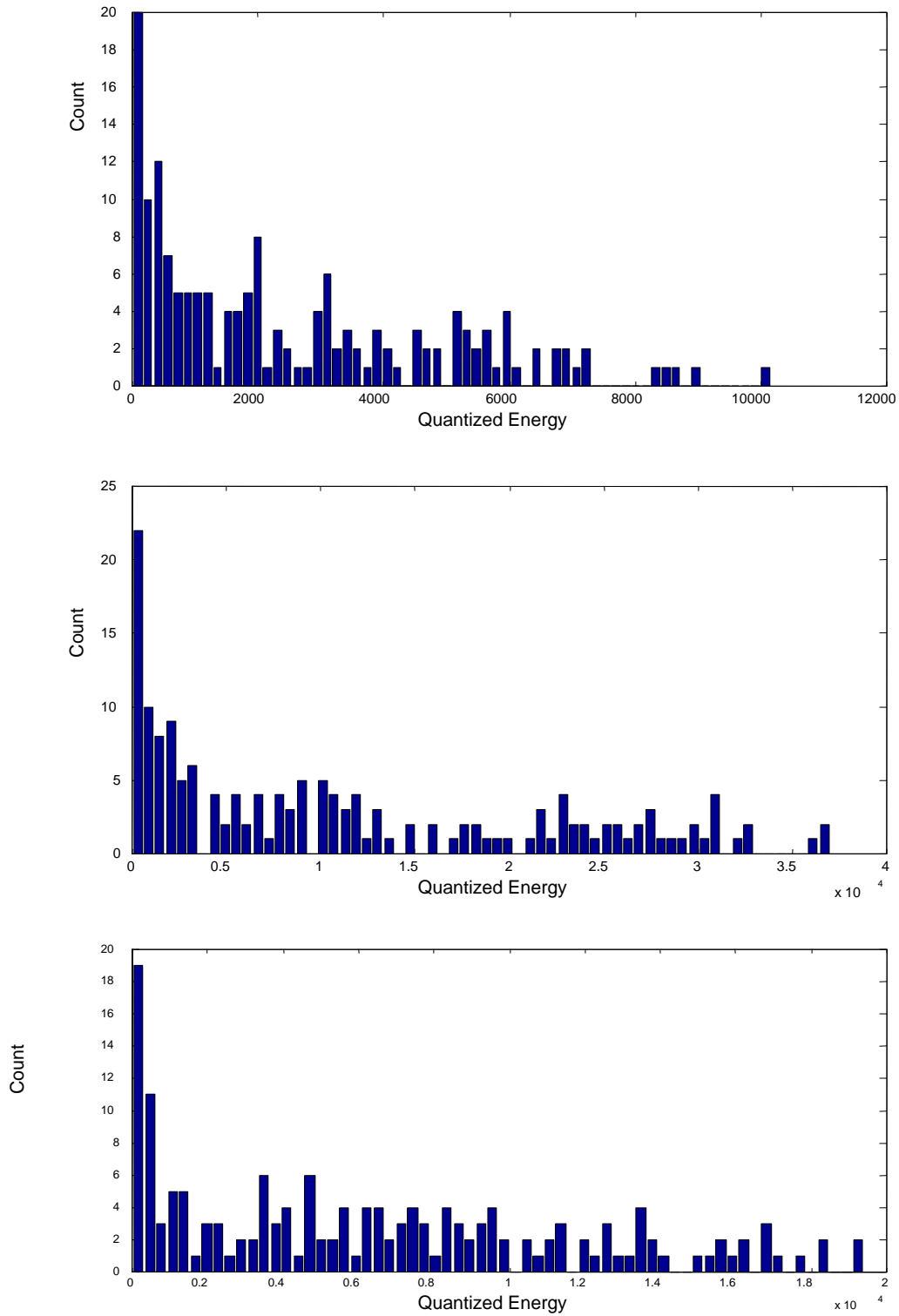


Figure 4.5: Energy distribution for channels 1, 2 and 3.

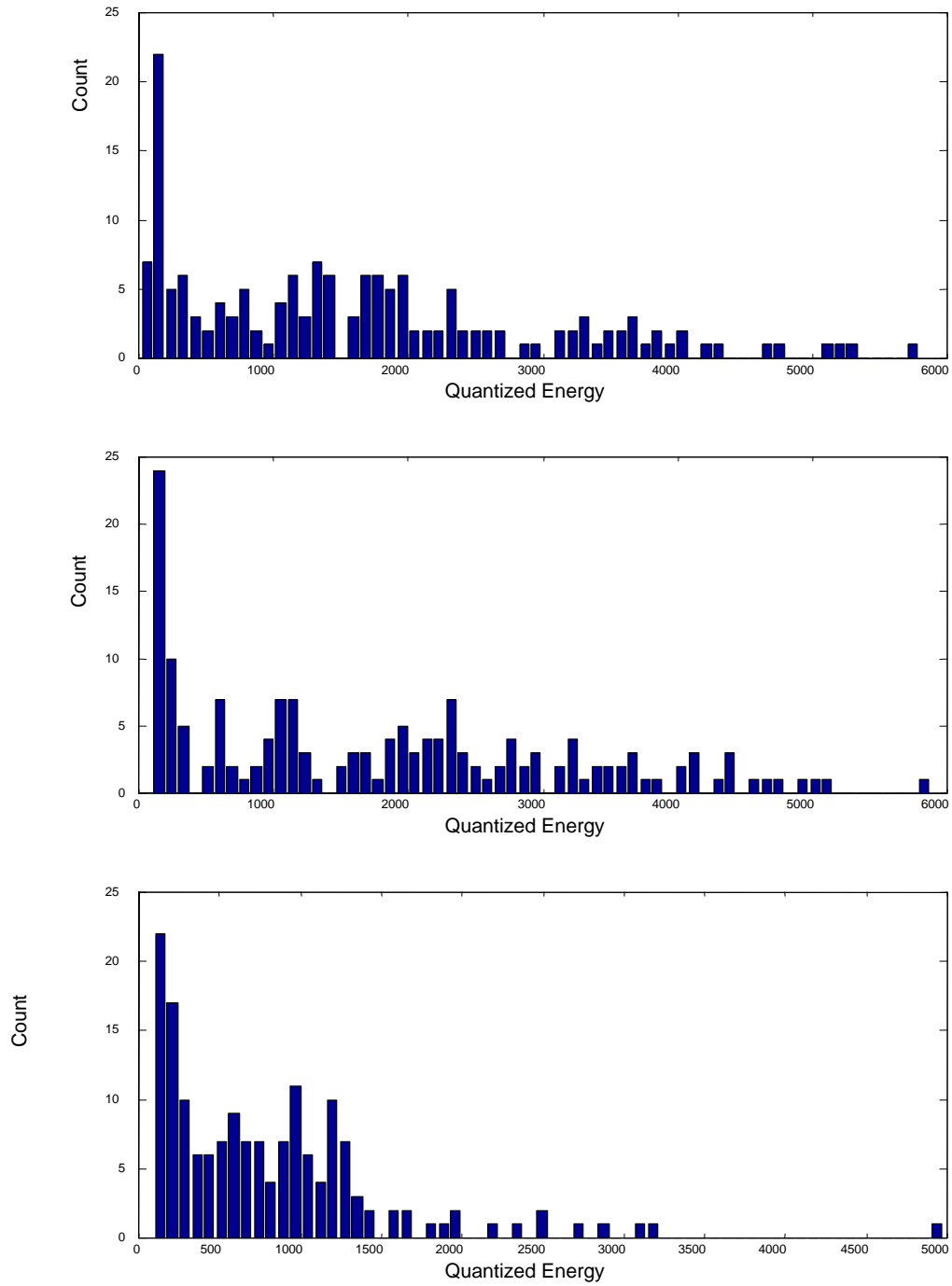


Figure 4.6: Energy distribution for channels 4, 5 and 6.

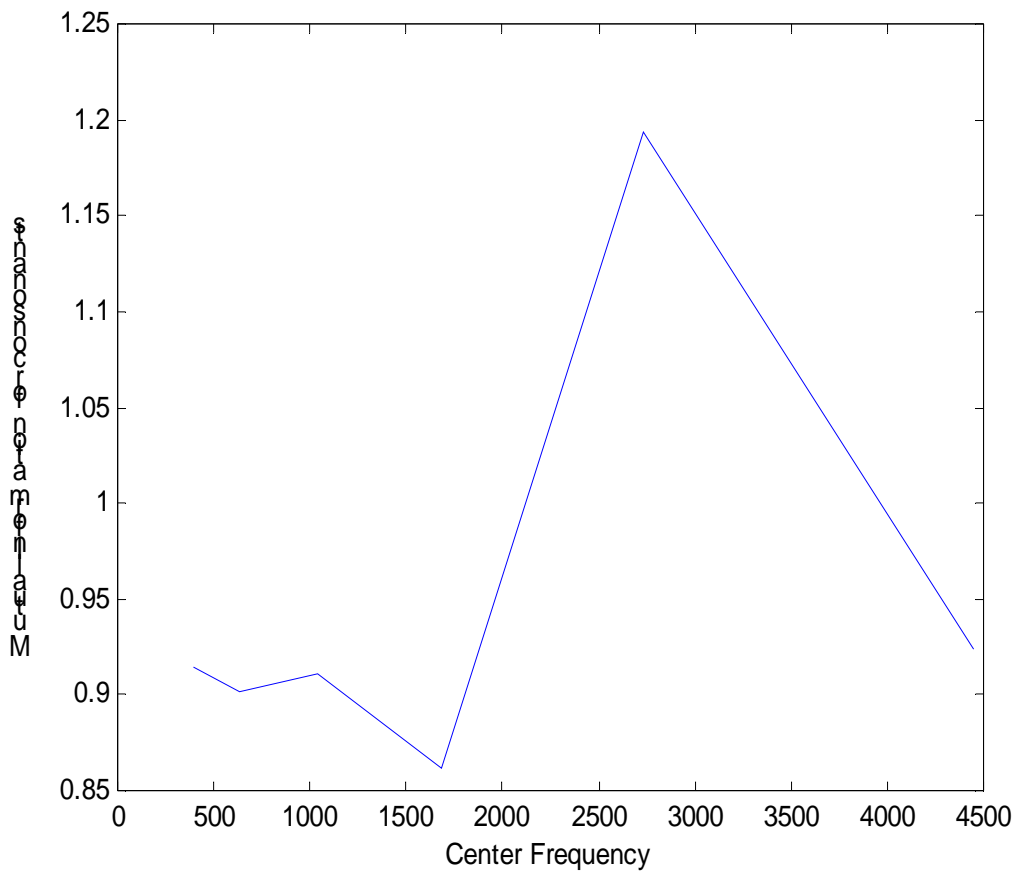


Figure 4.7: Mutual information between frequency bands and phonetic labels.

4.3.8 Results and Discussion

Here we present the results from the calculation of mutual information between spectral energy and phonetic labels for both the consonants and vowel stimuli.

Figure 4.8 presents the values of mutual information between the spectral energies and phonetic labels for the consonant stimuli. The results for vowel stimuli are presented in two plots, one for the vowels spoken by female speakers (Figure 4.9) and the other for the vowels spoken by male speakers (Figure 4.10).

The value of mutual information for a particular channel gives the amount of information shared between the phonetic labels and the energy in that channel, and thus gives the relative weight placed on that channel for phonetic classification. For both the consonants and vowels spoken by female speakers the plots of mutual information values are nearly flat. This indicates that information about phonetic classification is spread equally among all the channels.

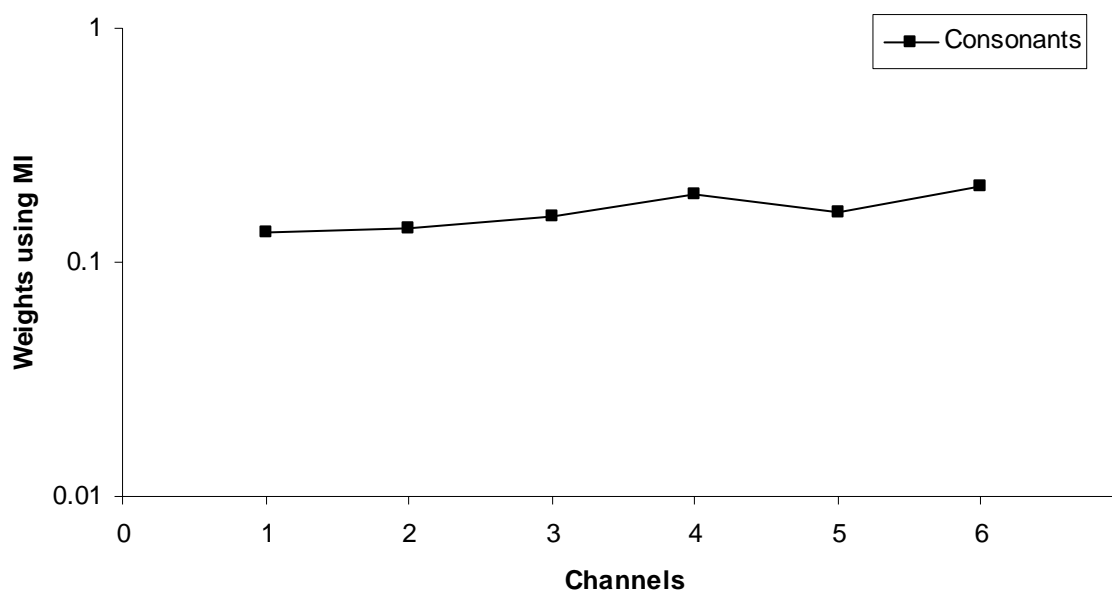


Figure 4.8: Mutual information between the spectral energy and the phonetic labels for consonant stimuli.

Also from Figure 4.8 we can note that channels 4, 5 and 6 contain relatively high amount of information about phonetic classification for consonant stimuli. This is consistent with the results from the least squares approach and reinforces the message that high frequency channels are more important for consonant identification.

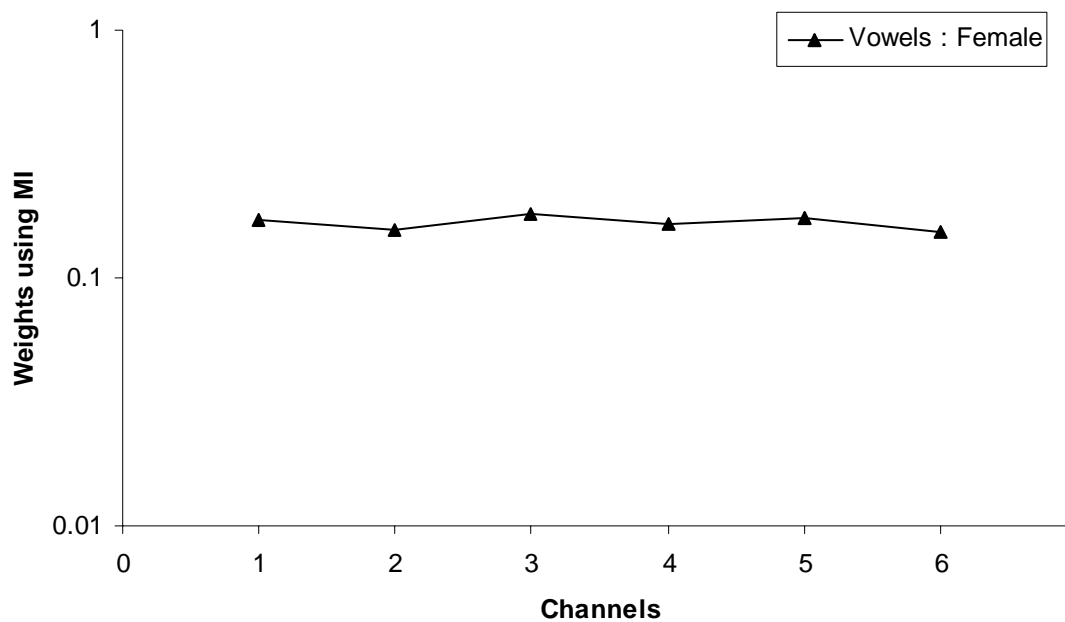


Figure 4.9: Mutual information between the spectral energy and the phonetic labels for vowel stimuli by female speakers.

From a close observation of Figure 4.9 we can see that the channels 1, 3 and 5 contain relatively high amount of information about the phonetic classification for vowel stimuli by female speakers. This is again consistent with the results from the least squares approach where channels 1, 3 and 4 were found to be most important. Channel 1 contains information about F1 which is vital for vowel identification and validates the relatively high weight for that channel. Again channels 3 and 5 code F2 information for female speakers, which is of primary importance for vowel identification and thus validate the relatively high weight for those channels.

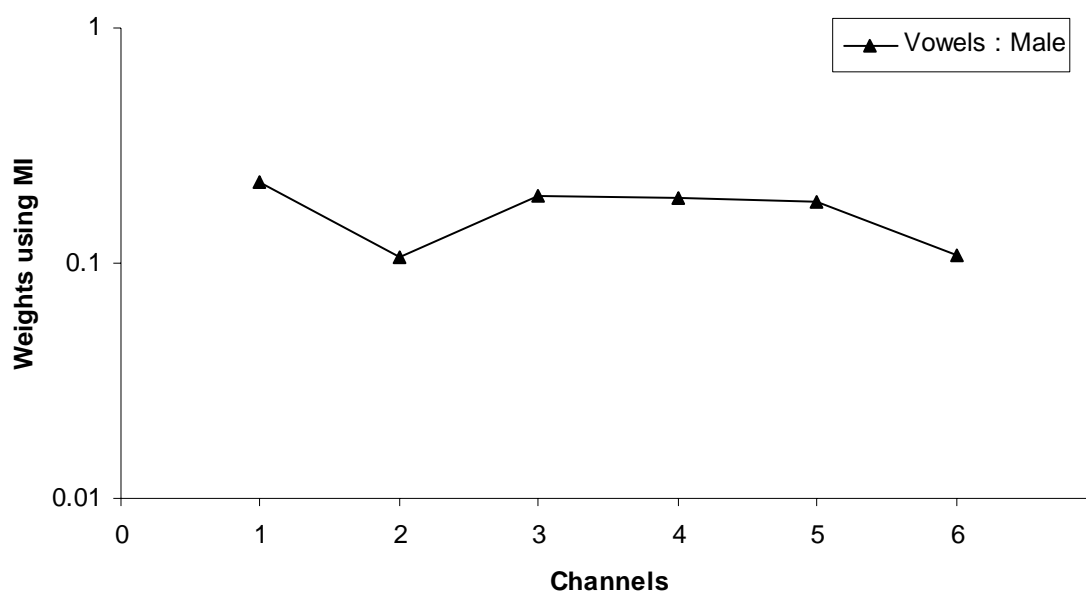


Figure 4.10: Mutual information between the spectral energy and the phonetic labels for vowel stimuli by male speakers.

From Figure 4.10 we can observe that there is notable variation among the weights for different channels for vowel stimuli by male speakers. The channels 1, 3 and 4 received the highest weight which is again consistent with the results from least squares approach. Channel 1, codes F1 information for male vowels and hence validates the relatively high weight for that channel. Channels 3 and 4 code the F2 information for male vowels and hence the relatively high weight for those channels.

Finally we compare the results from the calculation of mutual information with the weights obtained from the least squares approach in chapter 3. This comparison is very interesting since it gives an insight as to how the results from intelligibility tests performed compare with the predictions from theory using mutual information.

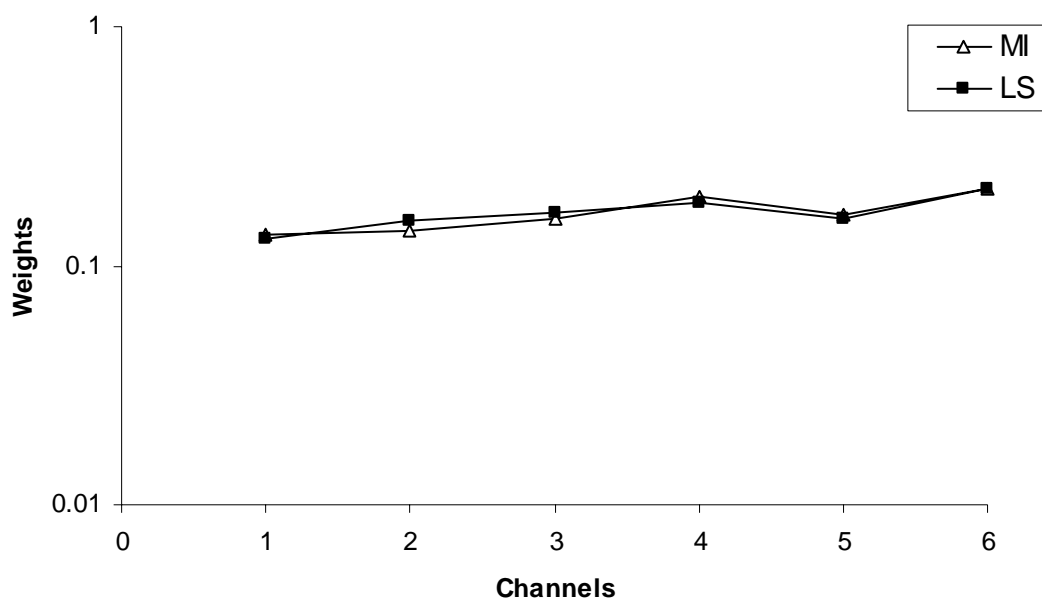


Figure 4.11: Comparison of weights obtained from mutual information and least squares methods for consonant stimuli.

In Figure 4.11 the weights obtained for consonants using mutual information and the least squares approach are plotted. The remarkable similarity between the weights calculated from both the approaches is evident. It must be noted that when calculating the mutual information for consonants the vowel /aa/ occurring before and after the consonant was ignored. This is consistent with the intelligibility tests where the subject knows beforehand that phoneme preceding and succeeding the unknown consonant is /aa/. Another interesting observation is the fact that the weights were notably different when the phoneme /aa/ was included in the computation of mutual information.

In Figures 4.12 and 4.13 the weights obtained from both the approaches are compared for vowel stimuli produced by female and male speakers respectively.

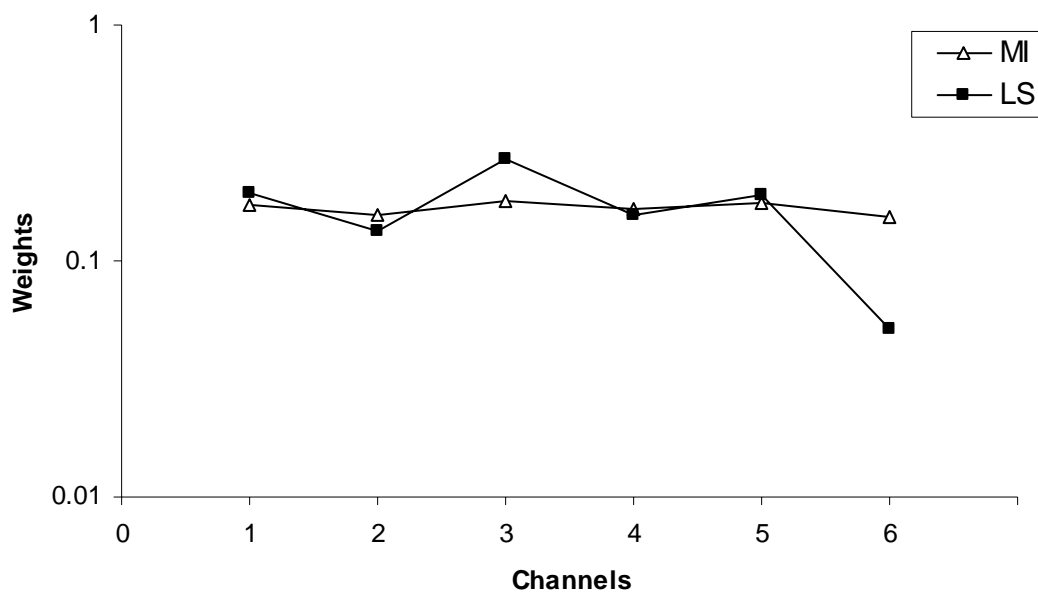


Figure 4.12: Comparison of weights obtained from mutual information and least squares methods for vowel stimuli by female speakers.

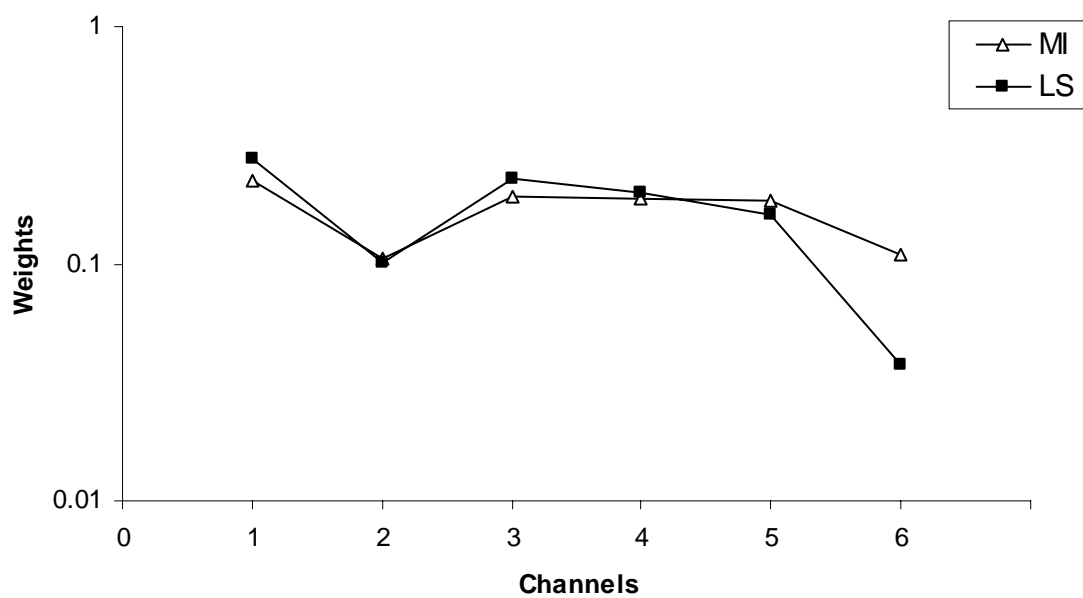


Figure 4.13: Comparison of weights obtained from mutual information and least squares methods for vowel stimuli by male speakers.

Again the similarity in weights obtained for vowel stimuli from both the approaches is evident. Also it can be noted that there is more variation in the weights obtained from the least squares approach compared to the weights obtained from mutual information. A possible explanation for this is variability in performance among the various listeners who participated in the intelligibility tests.

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

The current work focused on the intelligibility of speech with multiple disjoint bands in the spectrum. The present study addressed this question in a systematic manner considering all possible combinations of missing disjoint bands from the spectrum. Our study extended existing findings to show that high consonant recognition can be maintained even in the absence of disjoint frequency bands involving low, high and/or middle frequency information.

A novel approach for obtaining frequency-importance functions using a least squares criterion was derived. Also an information theoretic analysis was performed and the weights from least squares were shown to be in close agreement with the value of mutual information between frequency bands and the phonetic labels employed. The major conclusions from this study are summarized below:

- When a single “hole” was introduced in the spectrum, vowel and consonant recognition decreased. The degree of degradation in performance depended on the location of the “hole” or, equivalently, the frequency band removed. For vowels, there was a significant drop in performance when either of the frequency bands 1, 3 and 4 centered around 393, 1037 and 1685 Hz were removed. For consonants, there was a modest, yet significant, drop in performance when either of the frequency bands 4, 5 and 6, centered around 1685, 2736 and 4444 Hz were removed.

- Vowel recognition was affected the most, with the lowest performance (60% correct) obtained when channel 3, responsible for coding F2 information, was removed. Consonant recognition remained relatively high at around 70% correct even when high frequency channels were removed. Feature analysis indicated that the drop in consonant performance was primarily due to loss of “place” information. The manner and voicing features were not affected by the location of the “hole” in the spectrum.
- When two “holes” were introduced in the spectrum, vowel recognition decreased even further, and consonant recognition remained constant around 70% correct (the same as in the single “hole” condition).
- Vowel recognition performance was dependent on the frequency location of the pairs of bands removed. In particular, removing pairs of bands that contained F1 and/or F2 information caused a significant drop in performance.
- In contrast, consonant recognition was only mildly affected by the location of the pair of frequency bands removed. Consonant recognition remained robust at 70% correct, even when the middle and high frequency speech information was missing. This outcome is consistent with Lippmann’s [25] findings that accurate consonant recognition can be maintained even when the middle frequencies in the spectrum are absent. Our study extended Lippmann’s findings to show that high consonant recognition can be maintained even in the absence of disjoint frequency bands involving low, high and/or middle frequency information.

- The shapes of the frequency importance functions, derived using the least squares approach, were different for vowels and consonants. This is in agreement with the notion that different cues are used by listeners to identify consonants and vowels.
- For vowels, there was unequal weighting across the various channels. Channels 1, 3 and 4 received the largest weight. The frequency importance function for consonants was relatively flat, suggesting that all channels contributed equally to consonant identification, at least for the logarithmic filter spacing used in this study. This has important implications for cochlear implants. For CI listeners who are not able to extract useful information from *all* their electrodes, the logarithmic filter spacing might not be the optimal filter spacing.
- There is close agreement between weights or importance of frequency bands obtained from the least squares approach and the mutual information between the spectral energies and the phonetic labels.

Future work ought to be directed at studying the effect of various filter spacings and filtering schemes on the intelligibility of speech. Owing to the multitude of choices available, performing intelligibility tests with human subjects can be an onerous task. A better approach would be to use the information theoretic approach to evaluate the performance of various filtering strategies and then chose the best alternatives for further research consisting of intelligibility tests with human subjects.

APPENDIX A

FREQUENCY-IMPORTNACE FUNCTIONS FOR INDIVIDUAL SUBJECTS

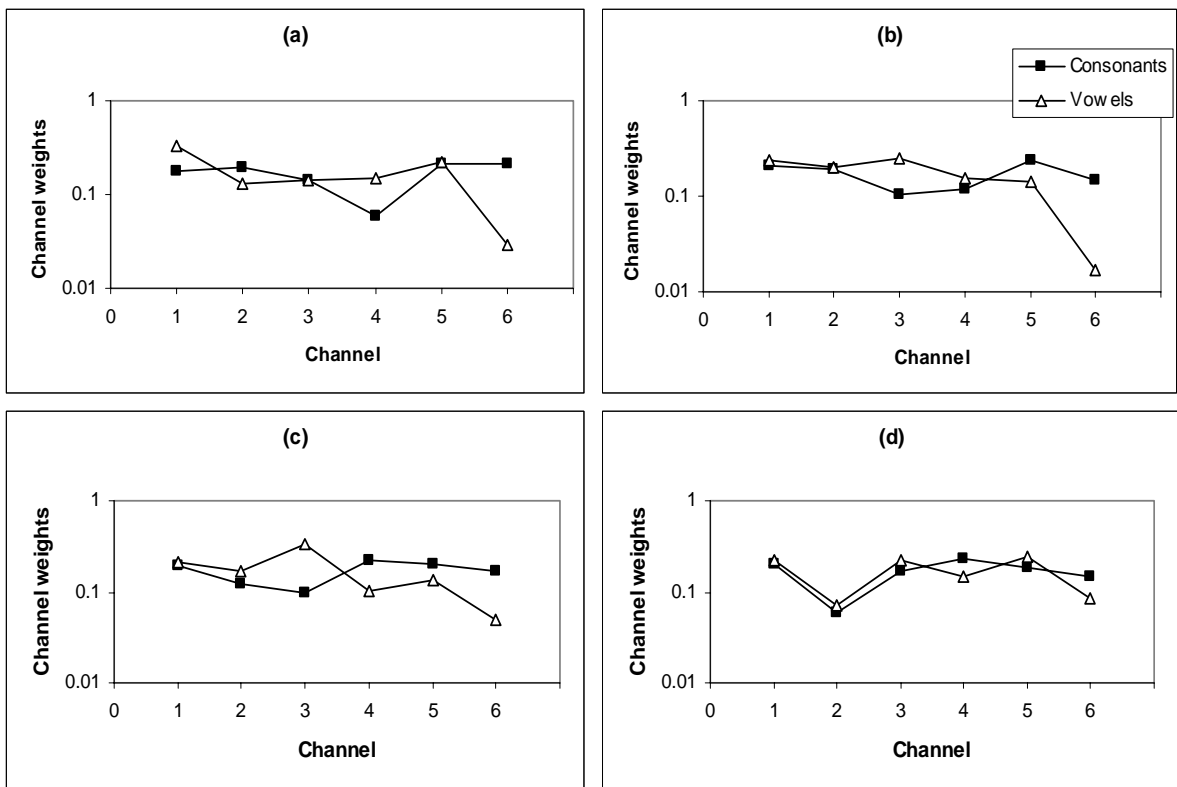


Figure A.1: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 1 and 2 respectively, panels (c) and (d) show the functions for subjects 3 and 4 respectively.

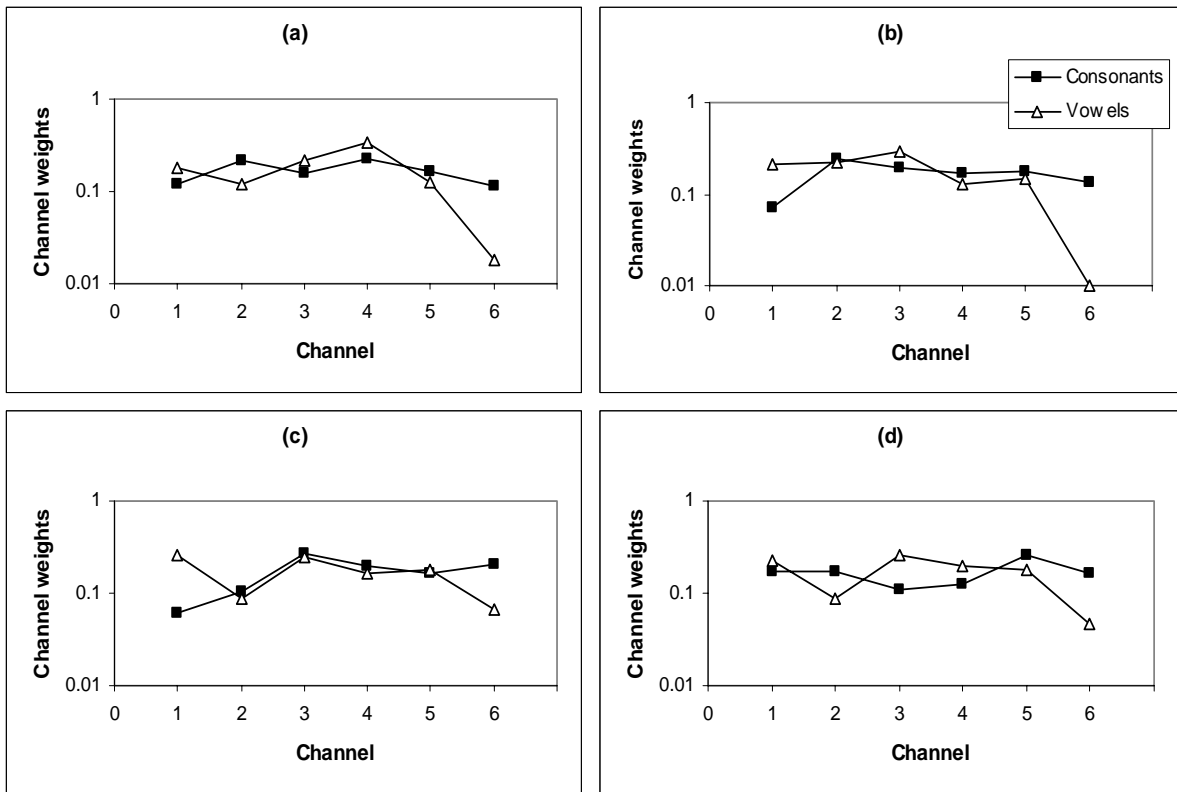


Figure A.2: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 5 and 6 respectively, panels (c) and (d) show the functions for subjects 7 and 8 respectively.

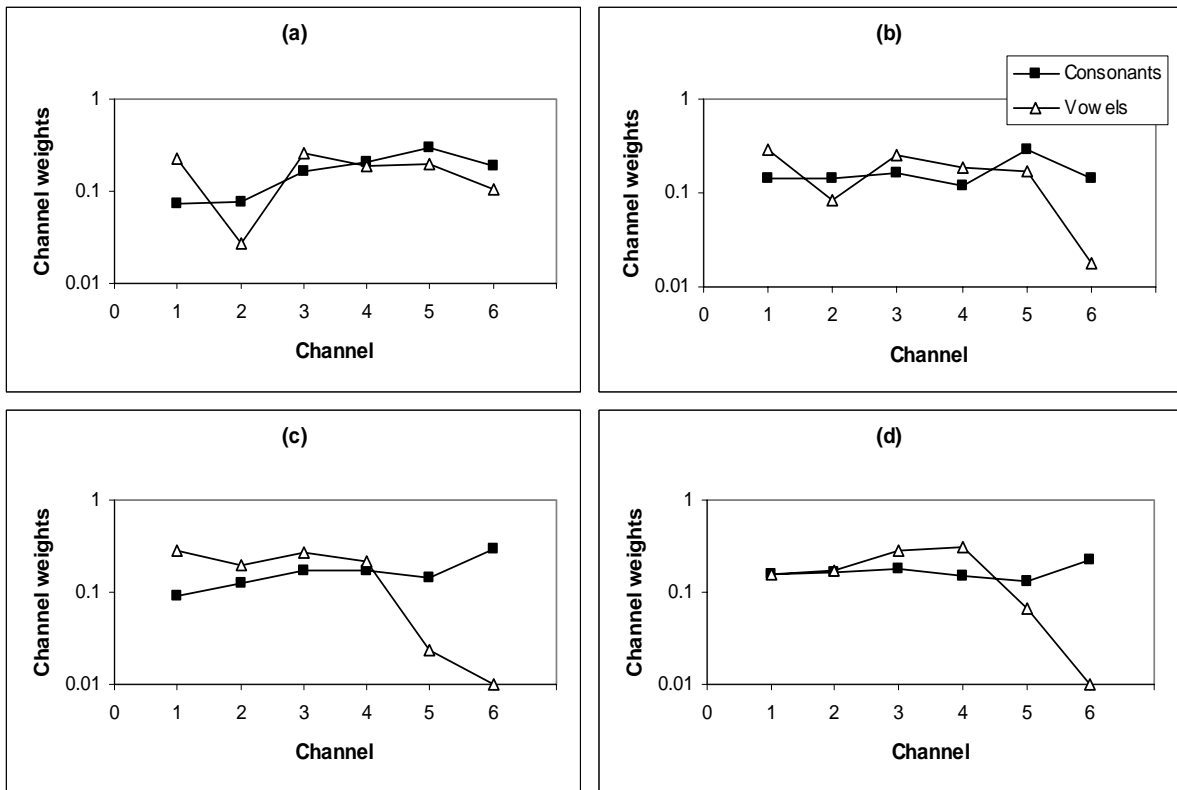


Figure A.3: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 9 and 10 respectively, panels (c) and (d) show the functions for subjects 11 and 12 respectively.

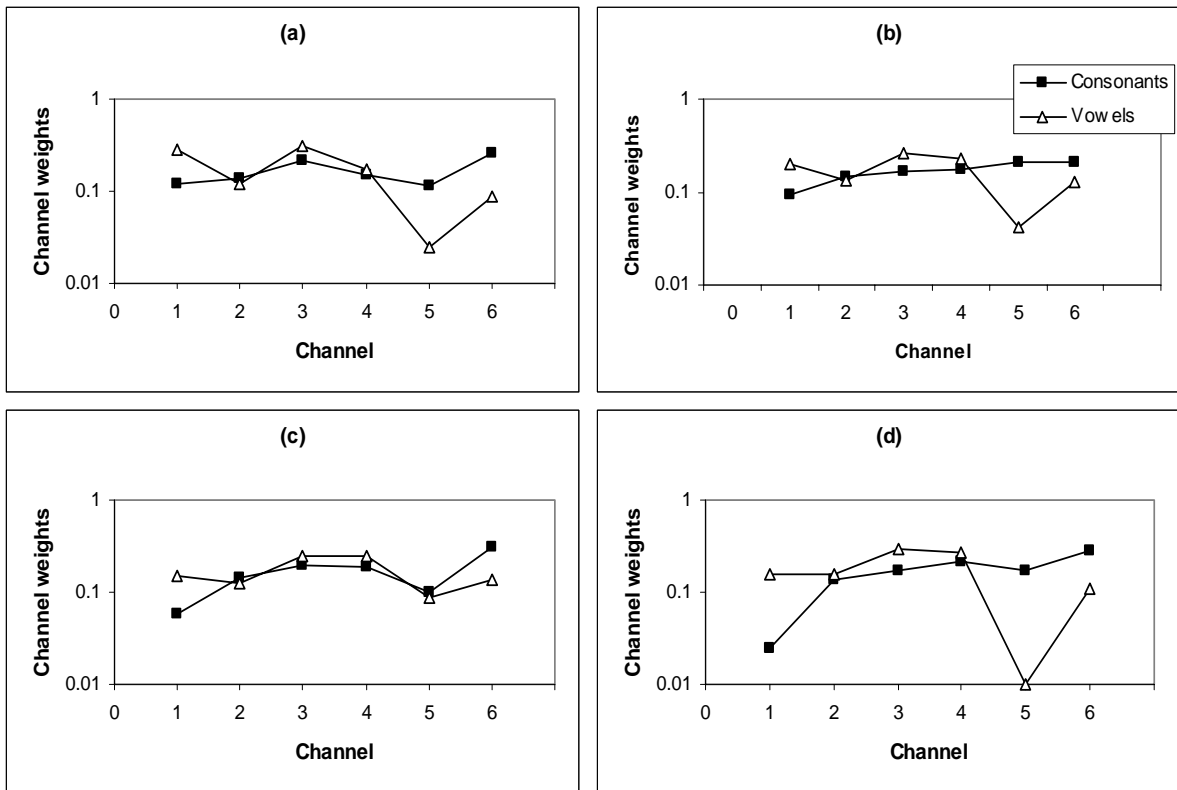


Figure A.4: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 13 and 14 respectively, panels (c) and (d) show the functions for subjects 15 and 16 respectively.

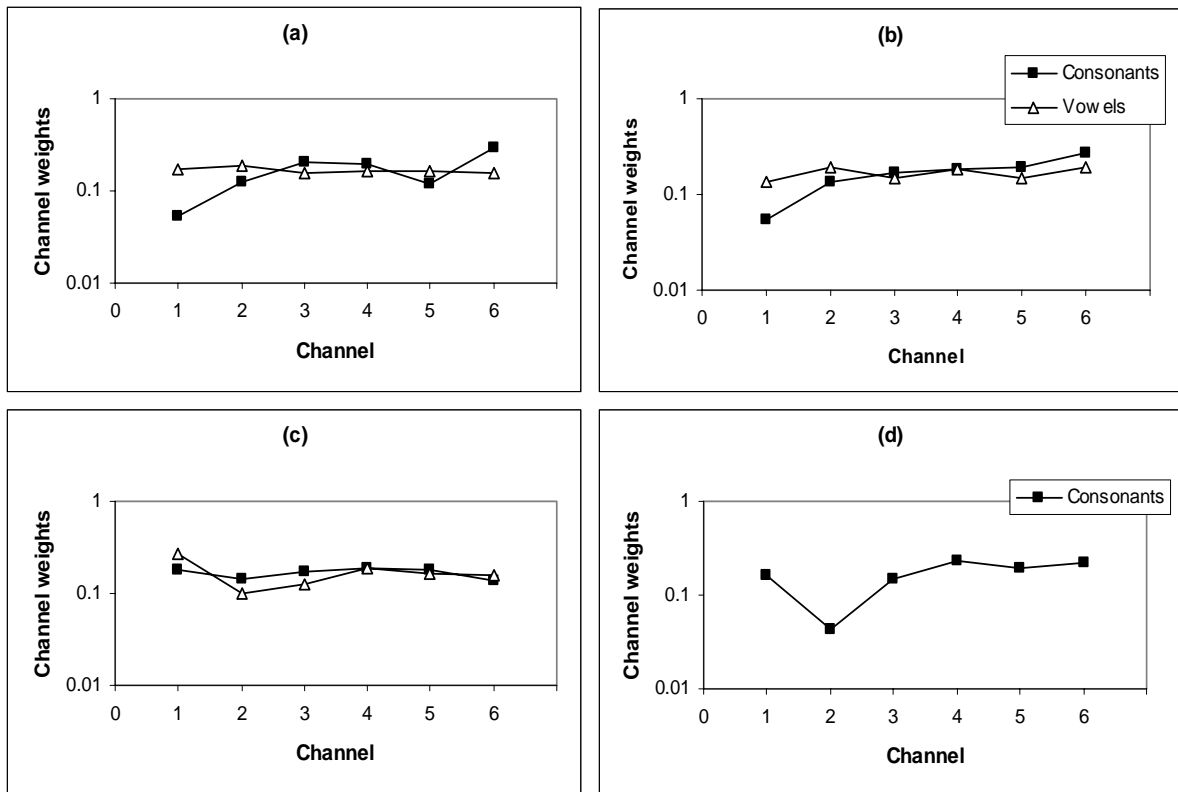


Figure A.5: Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for subjects 17 and 18 respectively, panels (c) and (d) show the functions for subjects 19 and 20 respectively.

APPENDIX B

PERCENT CORRECT RECOGNITION SCORES FOR VARIOUS CONDITIONS

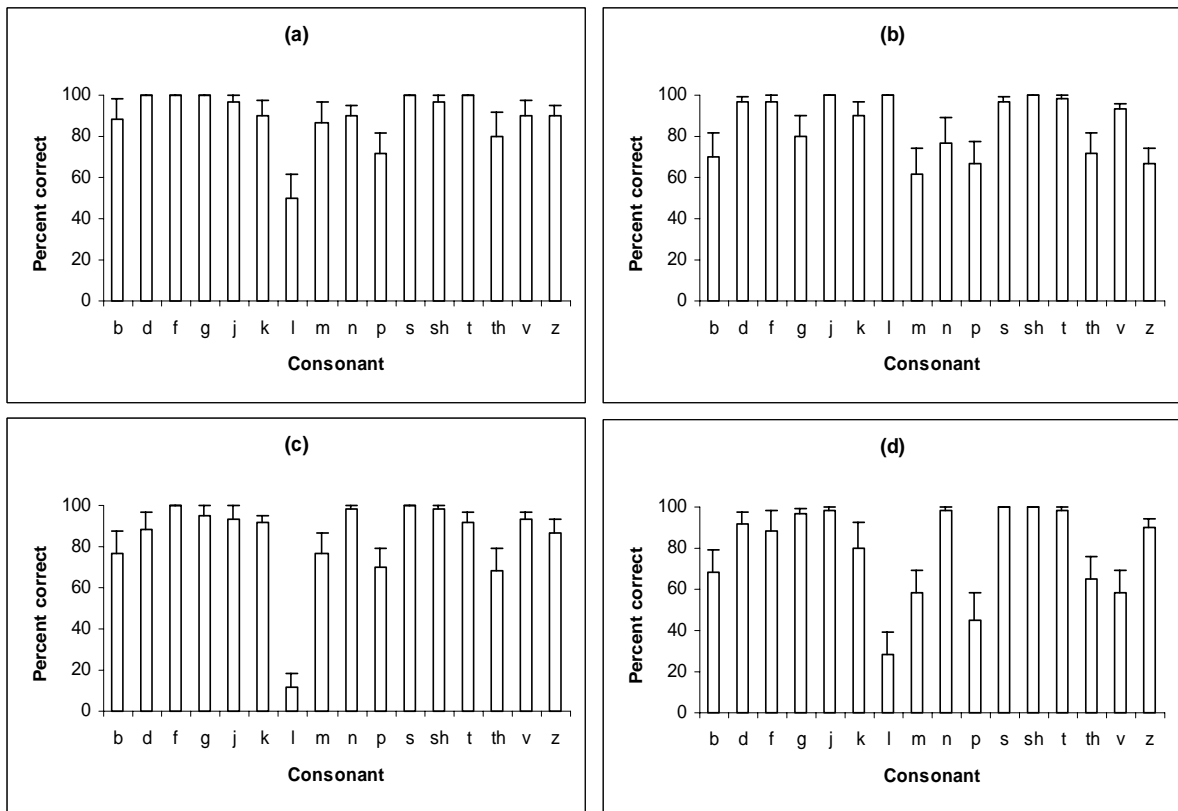


Figure B.1: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 0 and 1 respectively, panels (c) and (d) show mean percent correct scores for the conditions 2 and 3 respectively. Error bars indicate standard errors of the mean.

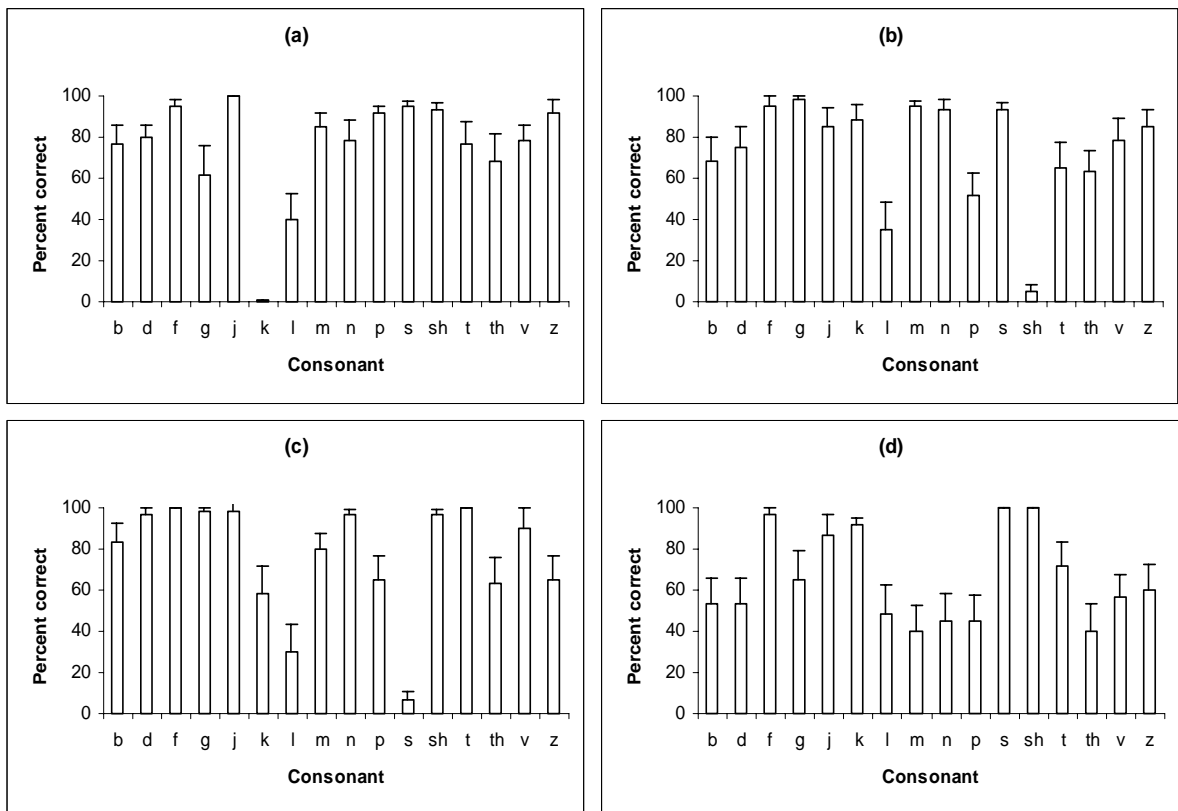


Figure B.2: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively, panels (c) and (d) show mean percent correct scores for the conditions 6 and 7 respectively. Error bars indicate standard errors of the mean.

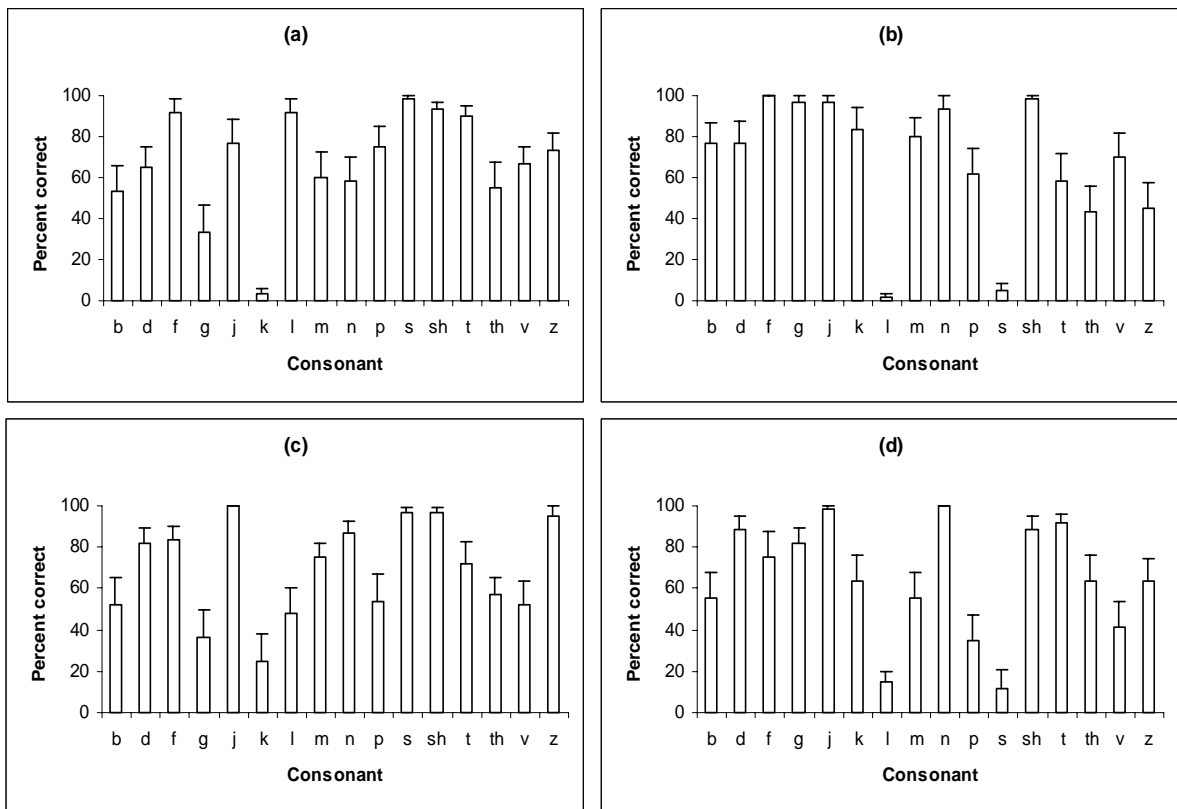


Figure B.3: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 9 and 15 respectively, panels (c) and (d) show mean percent correct scores for the conditions 16 and 18 respectively. Error bars indicate standard errors of the mean.

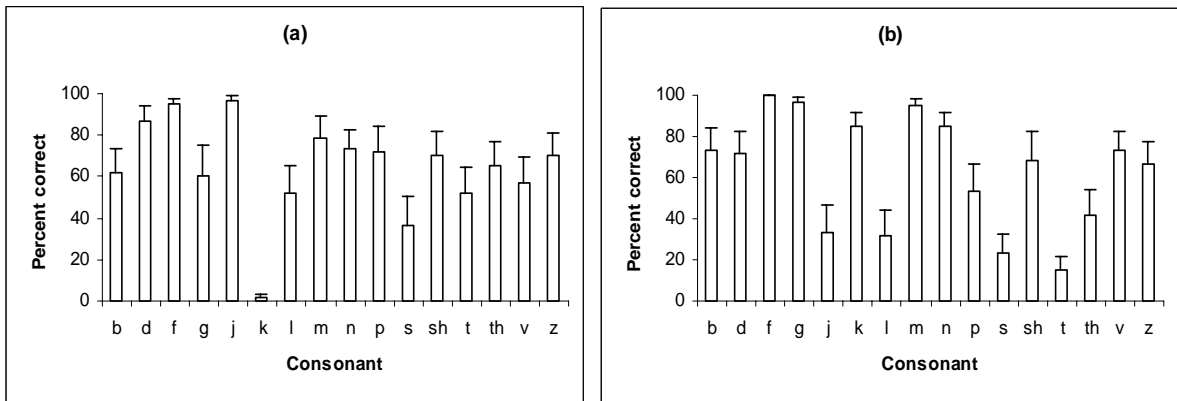


Figure B.4: Mean percent correct scores on individual consonant recognition. Panels (a) and (b) show mean percent correct scores for the conditions 20 and 21 respectively. Error bars indicate standard errors of the mean.

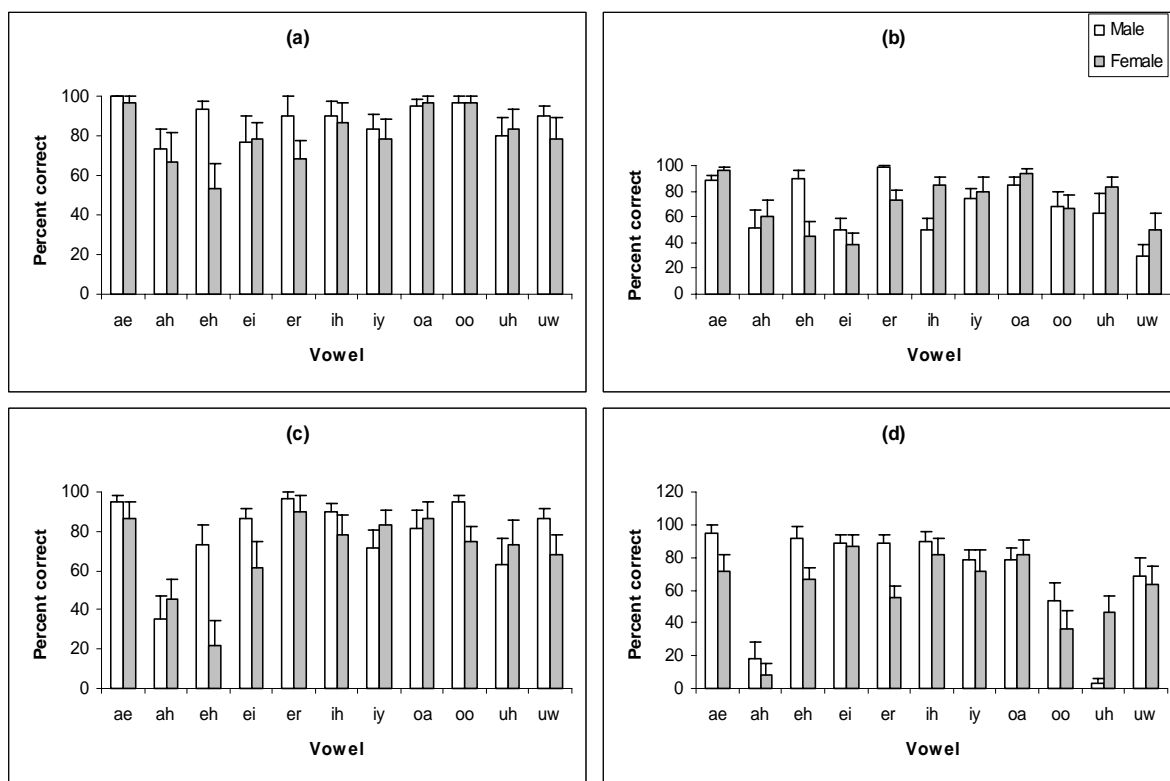


Figure B.5: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 0 and 1 respectively, panels (c) and (d) show mean percent correct scores for the conditions 2 and 3 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively.

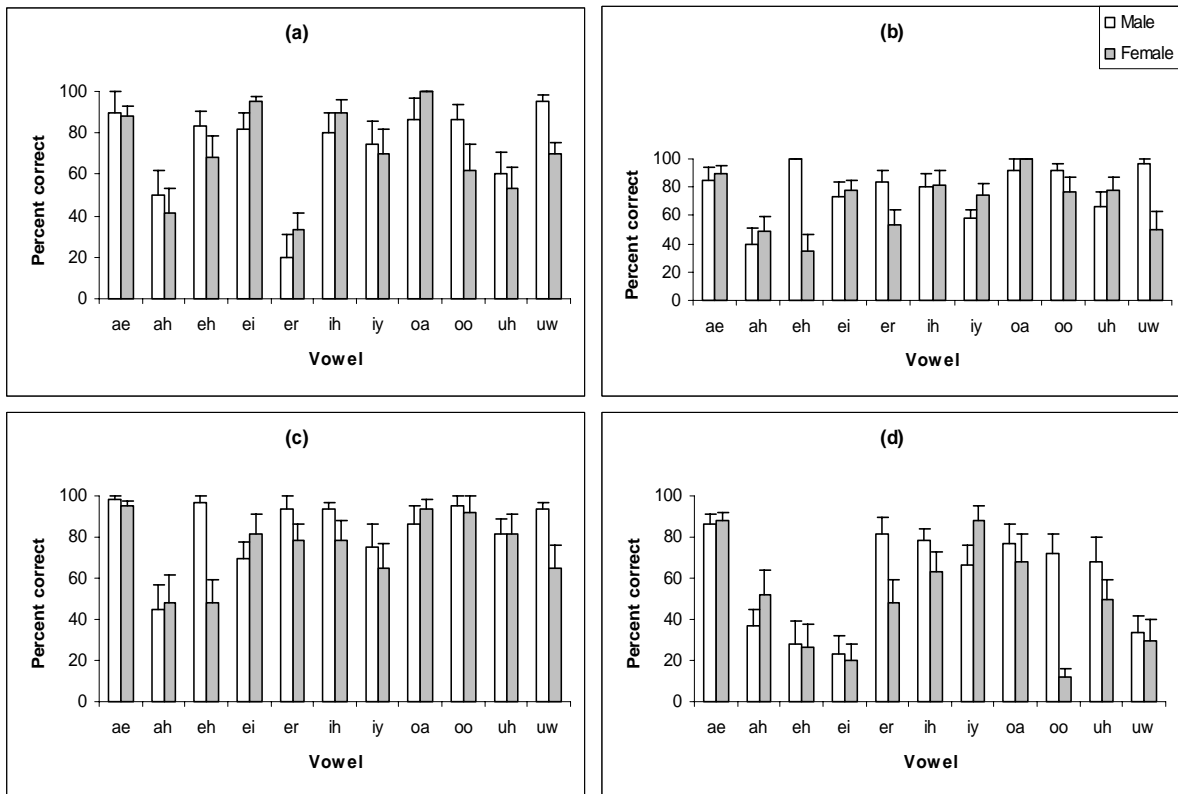


Figure B.6: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 4 and 5 respectively, panels (c) and (d) show mean percent correct scores for the conditions 6 and 7 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively.

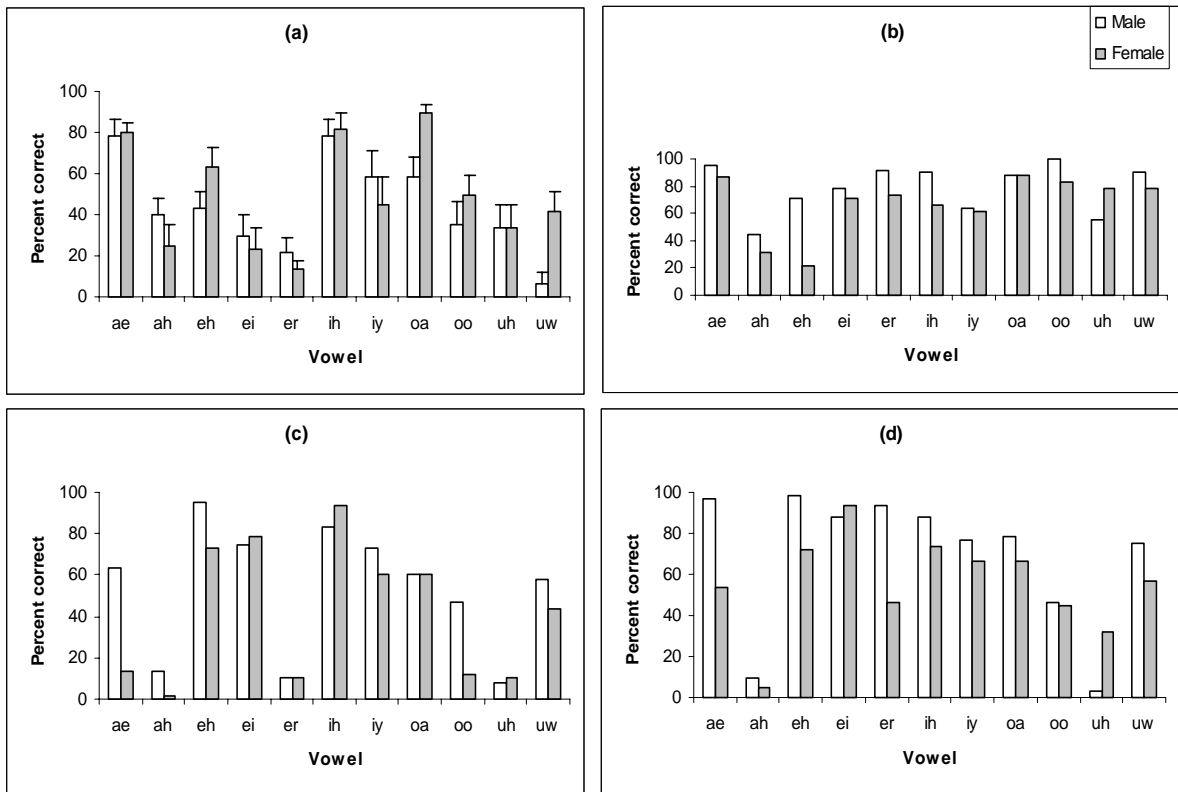


Figure B.7: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 9 and 15 respectively, panels (c) and (d) show mean percent correct scores for the conditions 16 and 18 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively.

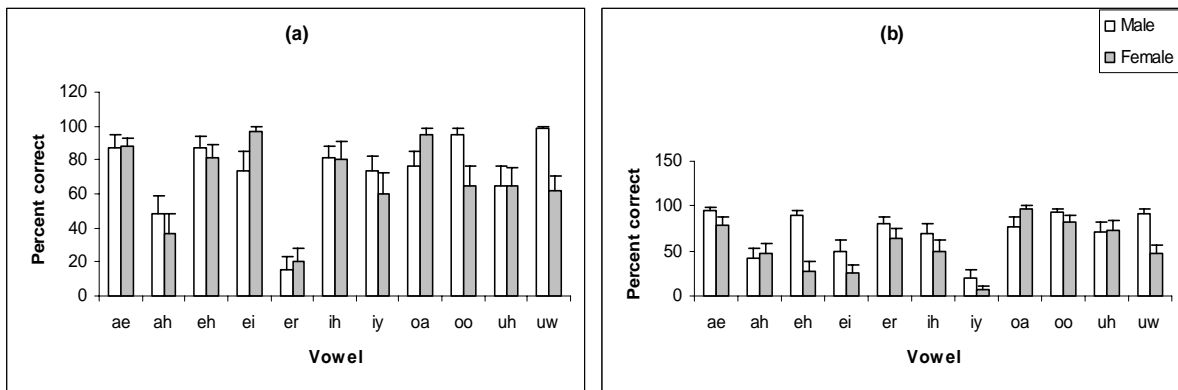


Figure B.8: Mean percent correct scores on individual vowel recognition. Panels (a) and (b) show mean percent correct scores for the conditions 20 and 21 respectively. Error bars indicate standard errors of the mean. The dark and white bars give the scores obtained with vowels produced by female and male speakers respectively.

REFERENCES

- [1] Ahumada, A., Jr., and Lovell, J., “Stimulus features in signal detection,” *J. Acoust. Soc. Am.*, vol. 49, pp. 1751 – 1756, 1970.
- [2] Allen, J. B., “How do humans process and recognize speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567 – 577, 1994.
- [3] ANSI (1969), ANSI S3.5 – 1969, “Specifications for audiometers,” ANSI New York.
- [4] Berg, B. G., “Analysis of weights in multiple observation tasks,” *J. Acoust. Soc. Am.*, vol. 86, pp. 1743 – 1746, 1989.
- [5] Bilmes, J. A., “Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling,” *ICASSP*, pp. 469 – 472, 1998.
- [6] Bilmes, J. A., “Joint distributional modeling with cross-correlation based features,” *Proc. of IEEE, ASRU*, Santa Barbara, 1997.
- [7] Breeuwer, M. and Plomp, R., “Speechreading supplemented with frequency-selective sound-pressure information,” *J. Acoust. Soc. Am.*, vol. 76, no. 3, pp 686 – 691, 1984.
- [8] Cover, T. M., and Thomas, J. A., *Elements of information theory*, Wiley, 1991.

- [9] Doherty, K. A., and Turner, C. W., "Use of a correlational method to estimate a listener's weighting function for speech," *J. Acoust. Soc. Am.*, vol. 100, pp. 3769 – 3773, 1996.
- [10] Dorman, M., Dankowski, K., McCandless, G. and Smith, L., "Consonant recognition as a function of the number of channels of stimulation by patients who use the Symbion cochlear implant," *Ear Hear.*, vol. 10, no. 5, pp. 288 – 291, 1989.
- [11] Dorman, M., Loizou, P., J. Fitzke, and Tu, Z., "The recognition of NU-6 words by cochlear implant patients and by normal-hearing subjects listening to NU-6 words processed in the manner of CIS and SPEAK strategies," *Annals of Otology, Rhinology and Laryngology*, vol. 109, no. 12, Suppl. 185, pp. 64 – 66, 2000.
- [12] Duggirala V., Studebaker G. A., Pavlovic C. V., and Sherbecoe R. L., "Frequency importance functions for a feature recognition test material," *J. Acoust. Soc. Am.*, vol. 83, no. 6, pp. 2372 – 2382, 1988.
- [13] Finkelstein, M., *Statistics at your finger tips*, Wadsworth, 1985.
- [14] Fishman KE, Shannon RV, Slattery WH., "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," *J Speech Hear Res*, vol. 40, no. 5, pp. 1201 – 1215, 1997.
- [15] Fletcher H., and Galt R. H., "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 89 – 151, 1950.
- [16] French, N. R., and Steinberg, J. C., "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90 – 119, 1947.

- [17] Friesen, L., Shannon, R., Baskent, D., and Wang, X., "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1150 – 1163, 2001.
- [18] Gallager, R. G., *Information theory and reliable communication*, John Wiley, 1968.
- [19] Grant, K. and Braida, L., "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2952 – 2960, 1991.
- [20] Haykin, S., *Adaptive filter theory*, Prentice Hall, 1996.
- [21] Hillenbrand, J., Getty, L., Clark, M. and Wheeler, K., "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, pp. 3099 – 3111, 1995.
- [22] Hirsh, I, Reynolds, E., and Joseph, M., "Intelligibility of different speech materials," *J. Acoust. Soc. Am.*, vol. 26, pp. 530 – 538, 1954.
- [23] Jenkins, G. M., and Watts, D. G., *Spectral Analysis and its applications*, Holden-Day, 1968.
- [24] Kryter, K., "Validation of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1698 – 1702, 1962.
- [25] Lippmann, R. P., "Accurate consonant perception without mid-frequency speech energy," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 66 – 69, 1996.
- [26] Loizou, P., "Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants," *IEEE Signal Processing Magazine*, vol. 15, no. 5, pp. 101 – 130, 1998.

- [27] Loizou, P., Dorman, M., and Tu, Z., "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 2097 – 2103, 1999.
- [28] Lutfi, R. A., "Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1333 – 1334, 1995.
- [29] Mehr, M. A., Turner, C. W., and Parkinson A., "Channel weights for speech recognition in cochlear implant users," *J. Acoust. Soc. Am.*, vol. 109, no. 1, pp. 359 – 366, 2000.
- [30] Miller, G. A., and Nicely, P. E., "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338 – 352, 1955.
- [31] Morris, A., Schwartz, J., and Escudier, P., "An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram," *Computer Speech and Language*, vol. 2, pp. 121 – 136, 1993.
- [32] Musch, H and Buus, S., "Using statistical decision theory to predict intelligibility. I. Model structure," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2896 – 2909, 2001.
- [33] Pollack, I., "Effects of high pass and low pass filtering on the intelligibility of speech in noise," *J. Acoust. Soc. Am.*, vol. 20, no. 3, pp. 259-266, 1948.
- [34] Richards, V. M., and Zhu, S. "Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients," *J. Acoust. Soc. Am.*, vol. 95, no. 1, pp. 423 – 434, 1994.
- [35] Riener, K., Warren, R. and Bahsford, J., "Novel findings concerning intelligibility of bandpass speech," *J. Acoust. Soc. Am.*, vol. 91, no. 4, pp. 2339, 1992.

- [36] Shannon, C. E., and Weaver, W., *The mathematical theory of communication*, Univ. of Illinois press, Urbana, 1971.
- [37] Shannon, R., Galvin, J., and Baskent, D., “Holes in hearing,” *J. Assoc. Res. Otolaryng.*, vol. 3, pp. 185 – 199, 2001.
- [38] Stark, H., and Woods, J., *Probability, Random Processes, and Estimation for Engineers*, Prentice-Hall, 1998.
- [39] Stickney, G., and Assmann, P., “Acoustic and linguistic factors in the perception on bandpass-filtered speech,” *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1157 – 1165, 2001.
- [40] Studebaker, G., Pavlovic, C., and Sherbecoe, R., “A frequency importance function for continuous discourse,” *J. Acoust. Soc. Am.*, vol. 81, no. 4, pp. 1130 – 1138, 1987.
- [41] Turner, C., Kwon, B., Tanaka, C., Knapp, J and Doherty, K., “Frequency importance functions for broadband speech as estimated by the correlational method,” *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1580 – 1585, 1998.
- [42] Tyler, R., Preece, J. and Lowder, M., The Iowa audiovisual speech perception laser videodisc. *Laser Videodisc and Laboratory Report*, Dept. of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City, 1987.
- [43] Warren, R. M., Riener, K. R., Bashford, J. A., Jr., and Brubaker, B. S., “Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits,” *Perception and Psychophysics*, vol. 57, pp. 175 – 182, 1995.

- [44] Yang, H., Vuuren, S., and Hermansky, H., "Relevancy of time-frequency features for phonetic classification measured by mutual information," *ICASSP*, vol. 1, pp. 225-228, 1999.
- [45] Zwolan, T., Collins, L., and Wakefield, G., "Electrode discrimination and speech recognition in postlingually deafened adult cochlear implant subjects," *J. Acoust. Soc. Am.*, vol. 102, no. 6, pp. 3673-3685, 1997.