

THE EFFECT OF NOISE ON THE SPECTRUM OF SPEECH

APPROVED BY SUPERVISORY COMMITTEE:

---

Dr. Philipos Loizou, Chair.

---

Dr. Andrea Fumagalli

---

Dr. Cyrus Cantrell

Copyright 2002

Gaurang Kishor Parikh

All Rights Reserved

*To my dear parents*

THE EFFECT OF NOISE ON THE SPECTRUM OF SPEECH

by

GAURANG KISHOR PARIKH, B.E.

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 TELECOMMUNICATION 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 on this 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. Fumagalli for obliging me to be on my defense committee and providing valuable suggestions and feedback.

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

I also thank my parents who sacrificed a lot for me and whose love, care and support helped me reach this stage in life.

Thanks are also due to all of my lab mates, from whom I learned a lot and whose companionship I enjoyed very much. My roommates also deserve appreciation, who provided me great inspiration at times.

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

# THE EFFECT OF NOISE ON THE SPECTRUM OF SPEECH

Gaurang Kishor Parikh, M.S.T.E.

The University of Texas at Dallas, 2002

Supervising Professor: Dr. Philipos C. Loizou

Real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. Little is known about the effect of noise on the spectrum of speech. Such knowledge could potentially help us develop better speech enhancement algorithms. This thesis investigates the affect of colored noise viz. multi-talker babble and speech-shaped noise on the spectrum of vowels and consonants. Multi-talker babble and speech-shaped noise were added to vowels and stop consonants at -5 to 15 dB SNR and the spectral effect of noise was quantified in terms of various acoustic measures: (a) spectral contrast of the noisy vowel spectra, (b) spectral distance between the noisy and clean vowel and consonant spectra for three frequency bands, (c) detection and estimation of first two formant frequencies in noise, (d) frequency deviation of first two formant frequencies in noise, (e) spectral tilt of stop consonants, and (f) burst frequency for the stop consonants.

Results showed that for vowels and stop consonants, the effect of colored noise on the frequency spectrum was non-uniform.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	vi
LIST OF FIGURES.....	xi
LIST OF TABLES.....	xiv
1. INTRODUCTION.....	1
2. LITERATURE REVIEW.....	3
2.1 Fundamental of speech production and classes of speech sound.....	3
2.1.1 Vowels.....	6
2.1.2 Consonants.....	11
2.2 Vowel perception.....	13
2.2.1 Formant frequencies.....	13
2.2.2 Vowel duration.....	15
2.2.3 Vowel fundamental frequency.....	16
2.3 Consonant perception.....	17
2.4 Spectral characteristics of vowels and consonants.....	20
2.4.1 Vowels.....	20
2.4.2 Consonants.....	22
3. ACOUSTIC MEASUREMENTS.....	24
3.1 Chapter outline.....	24

3.2	Speech material.....	24
3.2.1	Vowels.....	24
3.2.2	Consonants.....	25
3.3	Noise.....	25
3.3.1	Multi-talker babble.....	26
3.3.2	Speech-shaped noise.....	26
3.4	Segmentation.....	28
3.5	Spectral analysis.....	28
3.5.1	Filterbank approach.....	28
3.5.2	Critical band spacing.....	29
3.5.3	Logarithmic spacing.....	29
3.6	Acoustic measurements for vowels.....	32
3.6.1	Spectral contrast measurements.....	32
3.6.2	Spectral distance measurements.....	33
3.6.3	Formant frequency measurements.....	34
3.7	Acoustic measurements for consonants.....	36
3.7.1	Measurement of burst frequency.....	41
3.7.2	Measurement of tilt in the release burst spectrum.....	41
3.7.3	Measurement of spectral distance.....	44
4.	RESULTS.....	46
4.1	Chapter outline.....	46
4.2	Spectral contrast for vowels.....	46
4.2.1	Effect of number of channels on spectral contrast.....	47
4.2.2	Effect of type of noise on spectral contrast.....	47
4.3	Spectral distance for vowels.....	52
4.3.1	Effect of SNR on spectral distance.....	52
4.3.2	Effect of type of noise on spectral distance.....	52
4.4	Formant frequency measurements for vowels.....	56
4.5	Formant frequency deviation of vowels.....	60
4.6	Spectral distance measurement for consonants.....	61

4.7 Burst frequency measurement for stop consonants .....	65
4.8 Tilt measurement of burst frequency spectrum.....	65
5. SUMMARY AND CONCLUSIONS.....	71
REFERENCES.....	74
VITA	

## LIST OF FIGURES

Figure 2.1: A simplified tube model of human speech production system.....	4
Figure 2.2: Generation of voiced and unvoiced speech.....	5
Figure 2.3: Phonemes in American English.....	5
Figure 2.4: Time waveform and its corresponding frequency spectrum of speech .....	7
Figure 2.5: Spectrogram showing the first two formant frequencies.....	7
Figure 2.6: Frequency spectrum showing the first three formant frequencies of a vowel and its corresponding spectrogram.....	9
Figure 2.7: LPC plot of a vowel showing the first three formant frequencies.....	9
Figure 2.8: Peterson and Barney studies (1952). F1 and F2 frequencies computed for different vowels by them.....	10
Figure 2.9: Peterson and Barney studies (1952). Vowels plotted on F1-F2 space.....	10
Figure 2.10: Time waveform of a consonant and its corresponding LPC spectrum showing high energy in the high frequency range.....	12
Figure 2.11: Time waveform and its corresponding spectrogram showing the burst.....	12
Figure 3.1: (Top panel) PSD of multi-talker babble, (Bottom Panel) PSD of speech shaped noise.....	27
Figure 3.2: Plot of one sample frame (10 ms) of vowel /eh/. /. (Top panel) the 20-pole LPC spectrum. (Bottom panel) rms energy vs. frequency for a 21- channel filterbank. ....	32
Figure 3.3: LPC spectrum for a sample 10 ms frame of vowel /ae. Plot shows 22-pole spectrum of clean and noisy speech at 0 dB. The spectrum with solid line represents the clean speech spectrum and the one with dashed line represents the noisy speech spectrum .....	35
Figure 3.4: Examples of category 1 frame (Top panel) Multiple peaks in the proximity region. (Bottom panel) No peaks detected in F1 proximity.....	37
Figure 3.5: Examples of category 2 frame. (Top panel) No peaks detected in the proximity of F2. (Middle panel) Multiple peaks in the proximity of F2. (Bottom panel) Multiple peaks between F1 and F2 region .....	38
Figure 3.6: Examples of category 3 frame Neither F1 or F2 can be detected in these frames.....	39

Figure 3.7: Example of category 4 frame Both F1 and F2 can be reliably detected.....	40
Figure 3.8: Examples of /ba/ 20-order LPC spectrum used to find the burst frequency. Solid lines show burst frequency spectrum for quiet /ba/ consonant. Dashed lines show the /ba/ burst spectrum under different noise conditions. (Top panel) -5 dB speech shaped noise condition. (Second panel from top) 0 dB speech shaped noise condition. (Third panel from top) 5 dB speech shaped noise condition. (Last panel) 10 dB speech shaped noise condition.....	42
Figure 3.9: Plots of stop consonants derived using 20-pole LPC analysis. Panels (a) and (b) show plots of labials /p/ and /b/ respectively. Panels (c) and (d) show plots of alveolars /d/ and /k/ respectively. Panels (e) and (f) show plots of velars /g/ and /k/ respectively.....	43
Figure 4.1: Spectral contrast for individual vowels corrupted by speech-shaped noise. (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise. The vowels used on the x-axis are: iy as in heed, ih as in hid, ei as in hayed, eh as in head, ae as in had, oo as in hood, uh as in hud, uw as in who'd, ah as in hod, and er as in heard.....	48
Figure 4.2: Spectral contrast for individual vowels corrupted by multi-talker babble. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble. (Bottom panel) 10 dB multi-talker babble.....	49
Figure 4.3: Effect of number of channels on spectral contrast. (Top panel) Results for speech-shaped noise. (Bottom panel) Results for multi-talker babble.....	50
Figure 4.4: Comparison of spectral contrast for vowels corrupted by speech-shaped noise.....	51
Figure 4.5: Spectral distance results for individual vowels corrupted by speech-shaped. (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise.....	53
Figure 4.6: Spectral distance results for individual vowels corrupted by multi-talker. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble. (Bottom panel) 10 dB multi-talker babble.....	54
Figure 4.7: Average spectral distance for vowels. (Top panel) Spectral distance results for speech-shaped noise. (Bottom panel) Spectral distance results for multi-talker babble.....	55
Figure 4.8: Formants detected for speech-shaped noise. (Top panel) 0 dB. (Middle panel) 5 dB. (Bottom panel) 10 dB.....	57
Figure 4.9: Formants detected for multi-talker babble. (Top panel) 0 dB. (Middle panel) 5 dB. (Bottom panel) 10 dB.....	58
Figure 4.10: Formants detected for different SNR. (Top panel) Results for speech-shaped noise. (Bottom panel) Results for multi-talker babble.....	59

Figure 4.11: Spectral distance results for individual stop consonants corrupted by speech-shaped noise (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise.....	62
Figure 4.12: Spectral distance results for individual consonants corrupted by multi-talker babble. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble (Bottom panel) 10 dB multi-talker babble.....	63
Figure 4.13: Average spectral distance for stop consonants. (Top panel) Spectral distance results for speech-shaped noise. (Bottom panel) Spectral distance results for multi-talker babble.....	64
Figure 4.14: Burst frequency for clean stop consonants. (Top panel) Average value of clean frequency for stop consonants found out considering all the vowel contexts. (Bottom panel) Averaged burst frequency for a particular vowel context.....	66
Figure 4.15: Burst frequency for different difference between quiet and noisy consonants for various SNR for speech-shaped noise (Top panel) -5 dB SNR, with 0 dB SNR (Second panel from top), with 5 dB SNR (third panel from top) and 15 dB SNR (bottom panel).....	67
Figure 4.16: Burst frequency difference between quiet and noisy consonants for different SNR for multi-talker babble (Top panel) -5 dB SNR, with 0 dB SNR (Second panel from top), with 5 dB SNR (third panel from top) and 15 dB SNR (bottom panel).....	68
Figure 4.17: Comparison of speech-shaped and multi-talker babble results for burst frequency difference measurements.....	69
Figure 4.18: Measurement of tilt in the spectrum for different SNR. (Top panel) Slope measurements for different SNR for speech-shaped noise. (Bottom panel) Slope measurement for different SNR for multi-talker babble.....	70

## LIST OF TABLES

Table 3.1: The mean values of the first two formant frequencies (in Hz) of the male and female vowels used in this study.....	26
Table 3.2: Upper edge frequencies, lower edge frequencies, center frequencies, and bandwidths for 21-channel filterbank with critical band spacing.....	30
Table 3.3: Frequency spacing and bandwidths for 4, 6, 8, 12 channels with logarithmic spacing.....	31
Table 4.1: Mean $\Delta F$ values between formant frequencies of corrupted and clean vowels	61

# **CHAPTER ONE**

## **INTRODUCTION**

The problem of understanding the effect of noise on the spectrum of speech is very important in speech processing. Since the classic study by Peterson and Barney (1952) on the distribution of the vowel formant frequencies on the F1-F2 plane, many studies were reported on the perception of vowels (Strang, 1989), estimation of difference limens for formant discrimination (Hawks, 1994) and vowel modeling (Syrdal and Gopal, 1986). There were a few studies (Pickett, 1957; Nebalek, 1988; Nebalek and Dagenais, 1986) on vowel identification in noise, however those studies focused on identification errors and the relationship between vowel identification, hearing loss and age. Not many studies quantified the perceptual effect of noise. Little is known, for instance, how noise affects the different frequency bands of the spectrum.

Knowing how noise affects the spectrum of the speech is important for several reasons. For one, such knowledge could help us design better noise reduction algorithms that could potentially help hearing-impaired listeners' speech understanding in noise. Secondly, it can help us more generally to develop better speech enhancement algorithms. Quantifying the affect of noise on the speech by measuring various quality measures helps in design of better noise reduction algorithms by repeatedly optimizing those measures. In this thesis, we take the first step in quantifying the affect of multi-talker babble and speech-shaped noise on the spectrum of vowels and consonants. This thesis tries to answer several questions. Which

frequency band does noise affect the most, and which is the least? Or is the effect uniform across all frequencies? By what amount is the spectral contrast reduced? How are the two formant frequencies (F1 and F2), known to be major cues for vowel discrimination, affected? How does noise affect the burst spectrum (in terms of spectral tilt and burst frequency) of the stop consonants? The above questions are answered quantitatively by performing acoustic analysis on 11 vowels and 6 stop consonants embedded in -5 to 15 dB noise.

This thesis is organized as follows. Chapter 2 gives a literature review on vowel and consonant perception. Chapter 3 discusses the acoustic analysis performed on the vowels and consonants. Results and quantitative performance comparison is discussed in Chapter 4. Chapter 5 gives the conclusions and presents a summary of the work done and future work.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

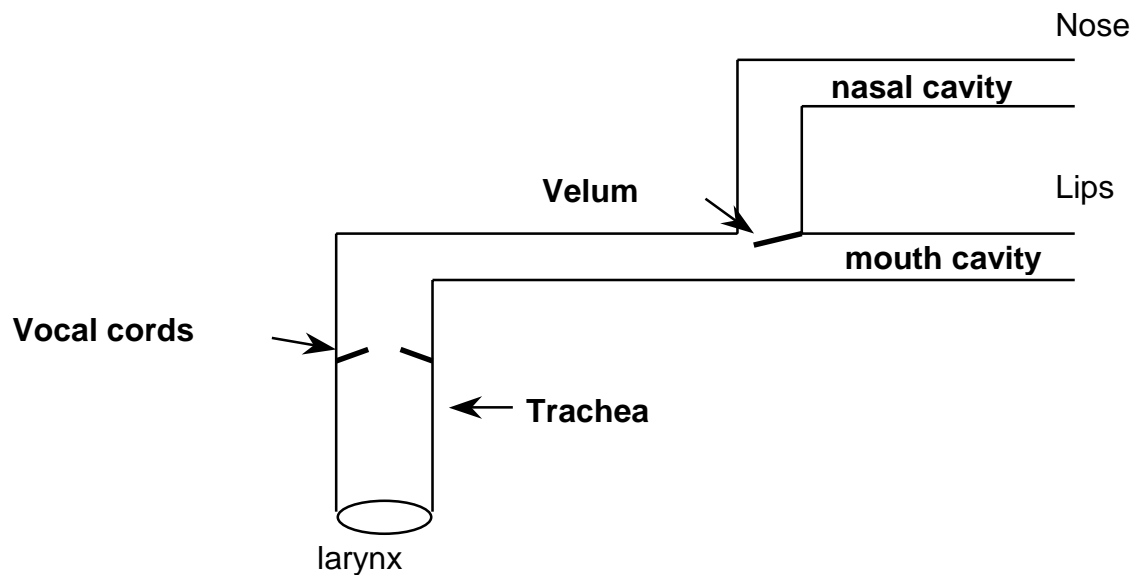
This chapter provides the literature review on the vowel and consonant perception. This chapter is divided as follows. Section 2.1 presents review on the production of speech and different classes of speech sound with emphasis on vowels and stop consonants. Section 2.2 and 2.3 discuss in detail about studies done in vowel and consonant perception. Finally, section 2.4 presents detailed explanation of the spectral characteristics of vowels and consonants.

#### **2.1 Fundamentals of speech production and classes of speech sounds**

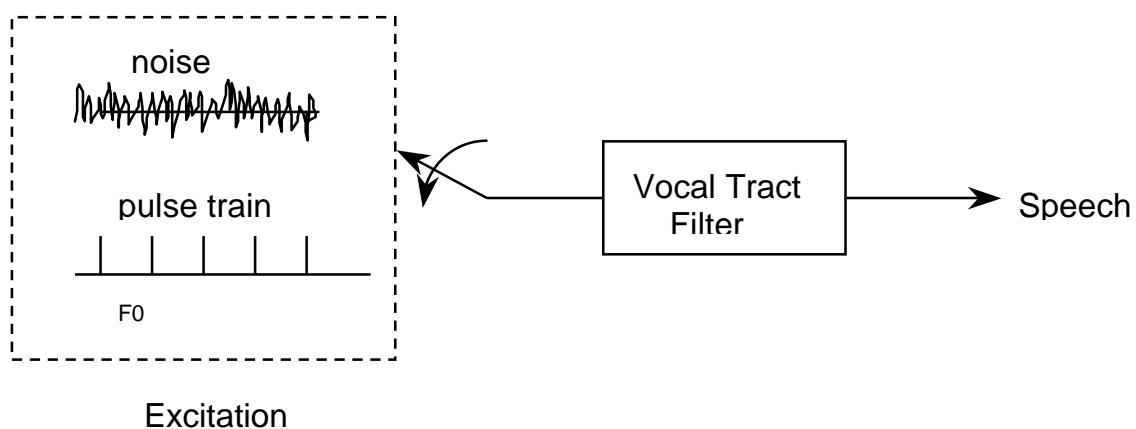
The speech signal consists of variation in pressure, measured directly in front of the mouth, as a function of time. The amplitude variations of such a signal correspond to deviations from atmospheric pressure caused by traveling waves. The signal is non-stationary and constantly changing as the muscles of the vocal tract contract and relax. Speech can be divided into sound segments which share some common acoustic properties with one another for a short interval of time. Sounds are typically divided into two broad classes: (a) vowels, which allow unrestricted airflow in the vocal tract, and (b) consonants, which restrict the airflow at some point and are weaker than vowels.

The human speech production mechanism consists of the lungs, trachea (windpipe), larynx, pharyngeal cavity (throat), buccal cavity (mouth), nasal cavity, velum (soft palate),

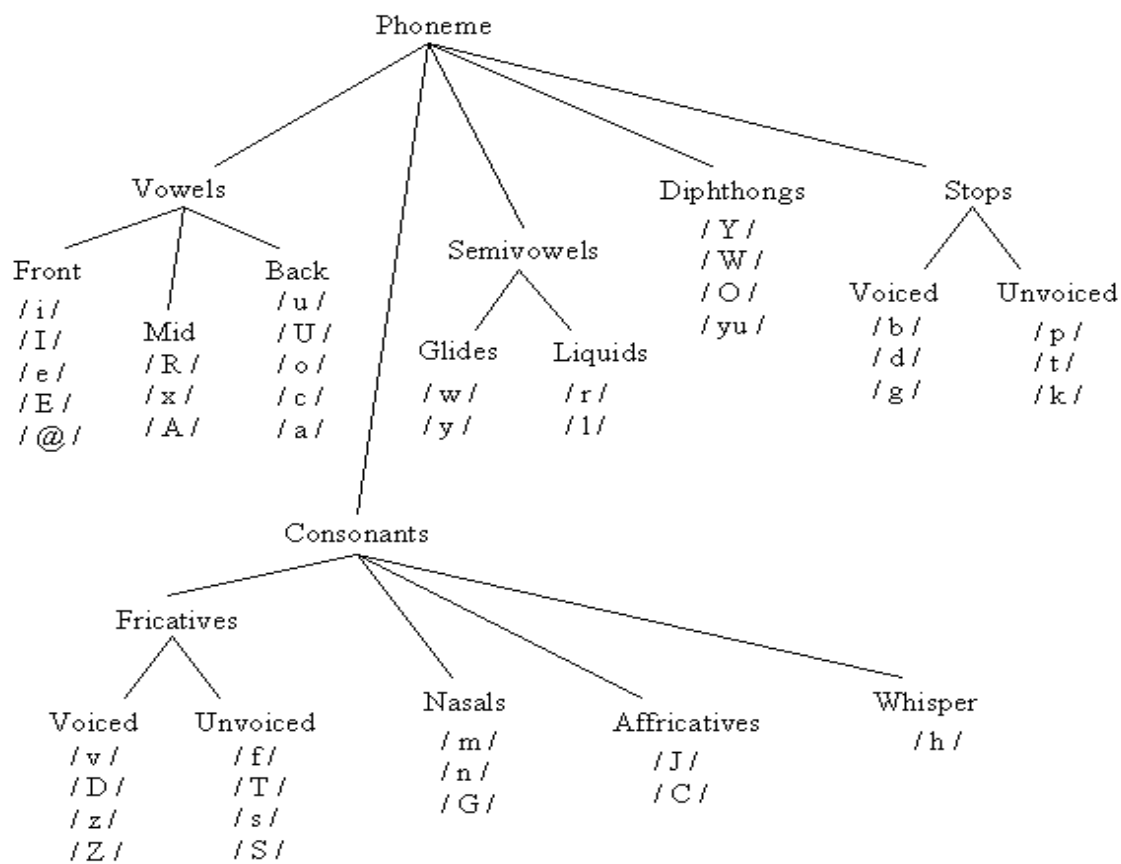
tongue, jaw, teeth and lips as shown in a simplified tube model in Figure 2.1. The lungs and trachea make up the respiratory subsystem of the mechanism. These provide the source of energy for speech when air is expelled from the lungs into the trachea. Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The source is periodic, resulting in voiced speech or aperiodic, resulting in unvoiced speech as shown in Figure 2.2. The voicing source occurs at the larynx at the base of the vocal tract, where airflow can be interrupted periodically by the vocal folds. The velum, tongue, jaw,



**Figure 2.1.** A simplified tube model of the human speech production system.



**Figure 2.2.** Generation of voiced and unvoiced speech.

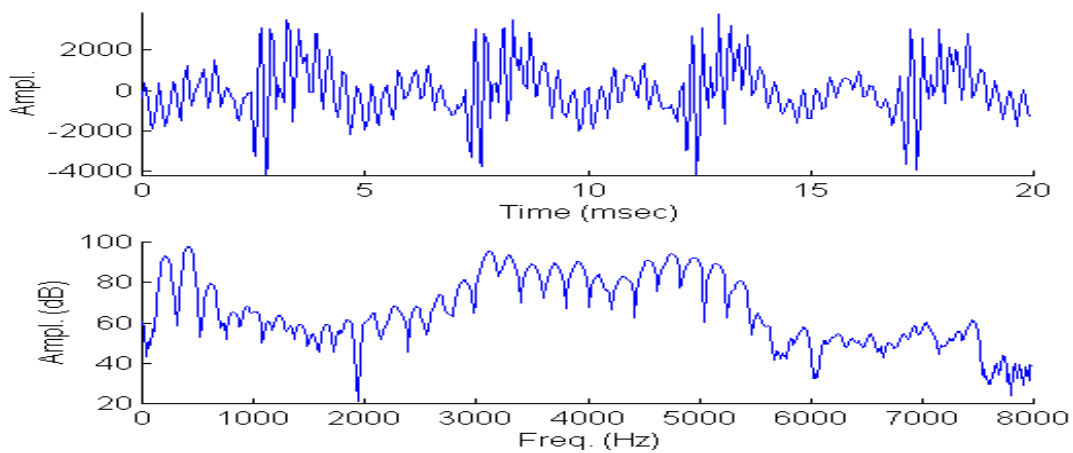


**Figure 2.3.** Phonemes in American English.

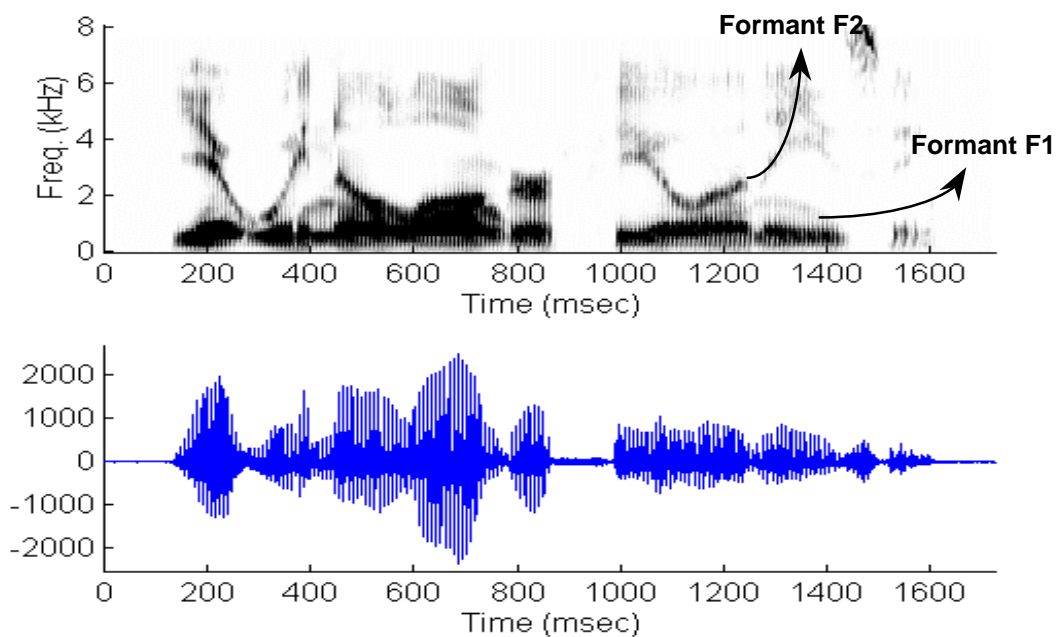
teeth and lips are known as the *articulators*. These provide the finer adjustments to generate speech. The excitation used to generate speech can be classified into *voiced, unvoiced, mixed, plosive, whisper* and *silence*. Any combination of one or more can be blended to produce a particular type of sound. A *phoneme* describes the linguistic meaning conveyed by a particular speech sound. The American English language consists of about 42 phonemes, which can be classified into vowels, semivowels, diphthongs and consonants (fricatives, nasals, affricatives and whisper) as shown in Figure 2.3.

### 2.1.1 Vowels

Vowels (including diaphthongs) are voiced, and have usually the largest amplitude among phonemes, and range in duration from 50 to 400 ms in normal speech. Figure 2.4 shows a brief portion of waveform for an English vowel and its corresponding frequency spectrum. Due to periodicity of the voiced excitation, the frequency spectrum exhibits harmonics with frequency spacing of  $F_0$  Hz where  $F_0$  is the *fundamental frequency* or the *pitch* of the vocal cord vibrations. Figure 2.5 shows a brief portion of waveform for an English vowel with its corresponding spectrogram. A spectrogram is a plot of frequency vs. time. The spectrogram reveals the amount of energy at different frequencies at different times. As seen from the spectrogram, the dark portions in the spectrogram represent the formant frequencies which are the dominant spectral peaks. The lower bold horizontal line is the first formant frequency ( $F_1$ ) and the upper dark portion represents the second formant frequency ( $F_2$ ). The formants can also be detected by inspection of the spectrum for dominant peaks as seen from Figure 2.6. The dominant peaks in the frequency spectrum can



**Figure 2.4.** Time waveform and its corresponding spectrum.

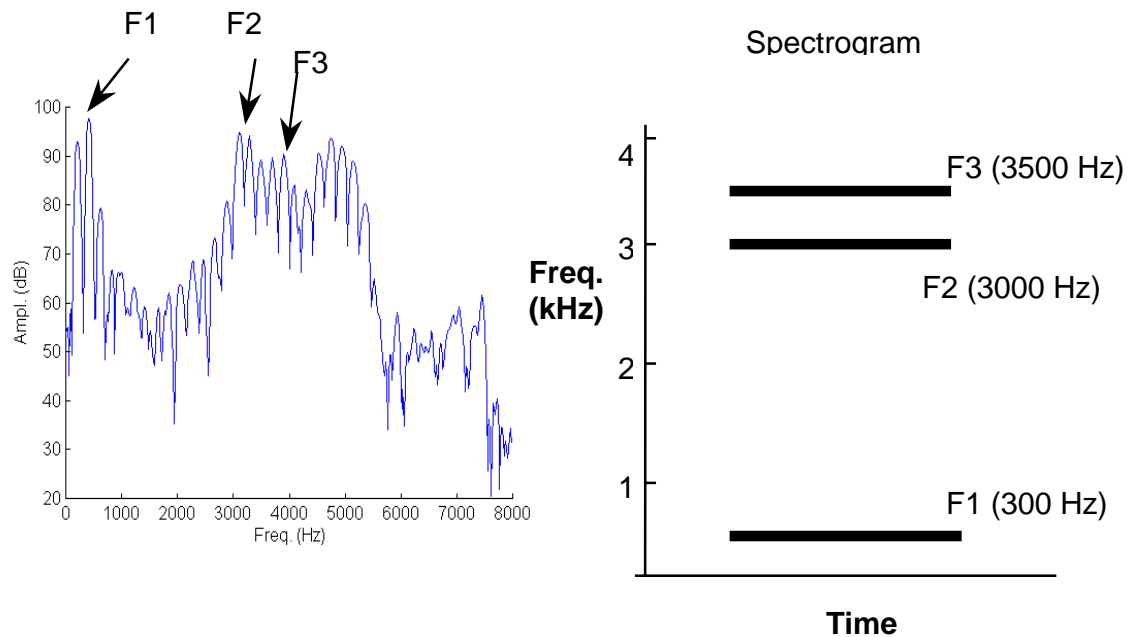


**Figure 2.5.** Spectrogram showing the first two formant frequencies.

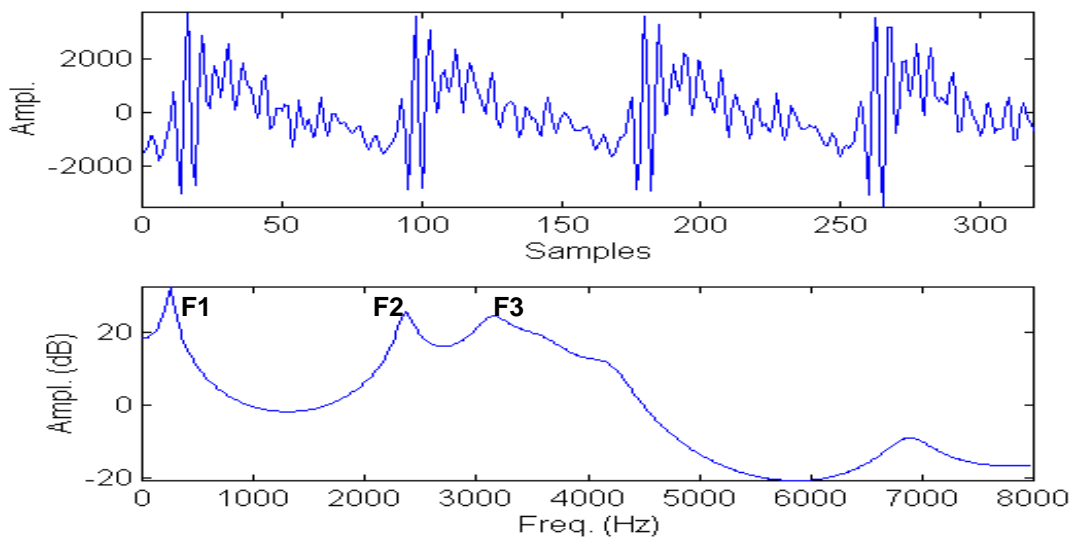
be detected as F1, F2 and F3 formant frequencies. The formant frequencies are normally derived from the Linear Prediction Coding (LPC) plot of the time waveform. Figure 2.7 shows the time waveform and the corresponding LPC plot with the formant frequencies F1, F2 and F3.

The time domain signal is not entirely periodic but is quasi-periodic as seen from Figure 2.7 due to the repeated excitations of the vocal tract by vocal fold closures. These excitations cause an abrupt signal increase once every period, after which amplitude of the signal decays exponentially with a time constant inversely proportional to the bandwidth of the formant(s) of the highest energy (F1). F1 can be readily identified in the time plots of many vowels as the inverse of the period of the dominant oscillation within a pitch period.

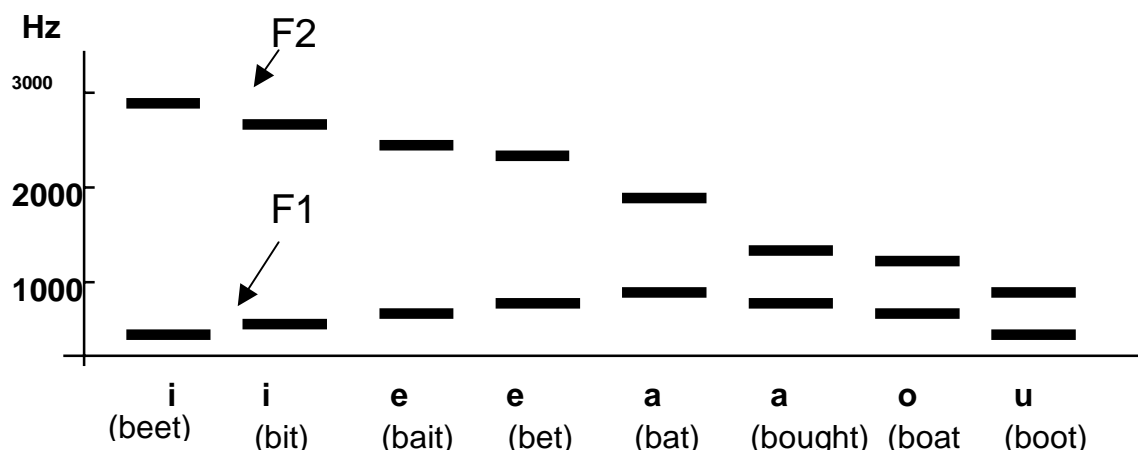
The first three formant frequencies are important as the vowels are distinguished primarily by the locations of their first three formant frequencies (Delattre *et al.*, 1952). In a study at Haskins Laboratories, Delattre *et al.* produced vowels by synthesizing steady state formants on the Pattern Playback. Pattern Playback was a machine developed at Haskins Laboratories which enabled the experimenter to draw the acoustic pattern, and synthesize the acoustic pattern. Figure 2.9 shows the vowels synthesized using the Pattern Playback machine. The results of this study were later reaffirmed by studies done by Fry *et al.* (1962) in which the vowels synthesized using the formant frequencies extracted from the natural vowels, produced satisfactory results. Peterson and Barney (1952) computed the mean formant values of ten different vowels using 32 male and 28 female speakers as shown in Figure 2.9. The speakers repeated twice each ten words of the form /hVd/, where V was one of the ten vowels. As seen from the figure there is a substantial overlap in the F1-F2 frequencies of the



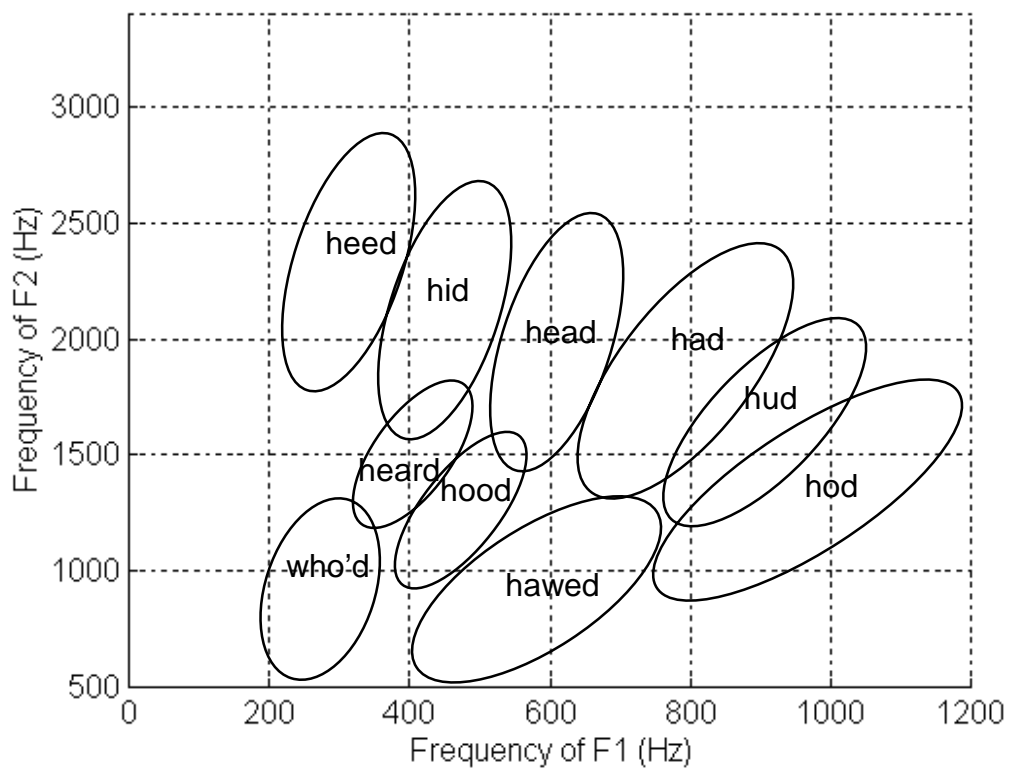
**Figure 2.6.** Spectrum showing the first three formant frequencies of a vowel and its corresponding spectrogram.



**Figure 2.7.** LPC plot of a vowel showing the first three formant frequencies.



**Figure 2.8.** Vowels synthesized using the above F1 and F2 frequencies (Delattre *et al*, 1952).



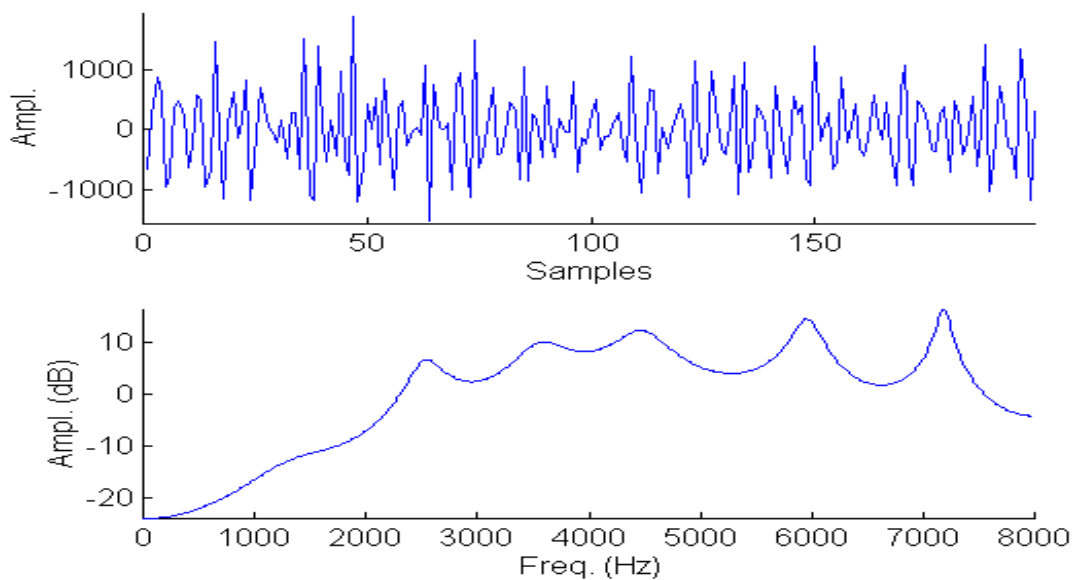
**Figure 2.9.** Vowels plotted on F1-F2 space (Peterson and Barney, 1952).

same vowel when spoken by different speakers. One of the reasons could be the varying length of the vocal tract since the formants differ considerably for different speakers.

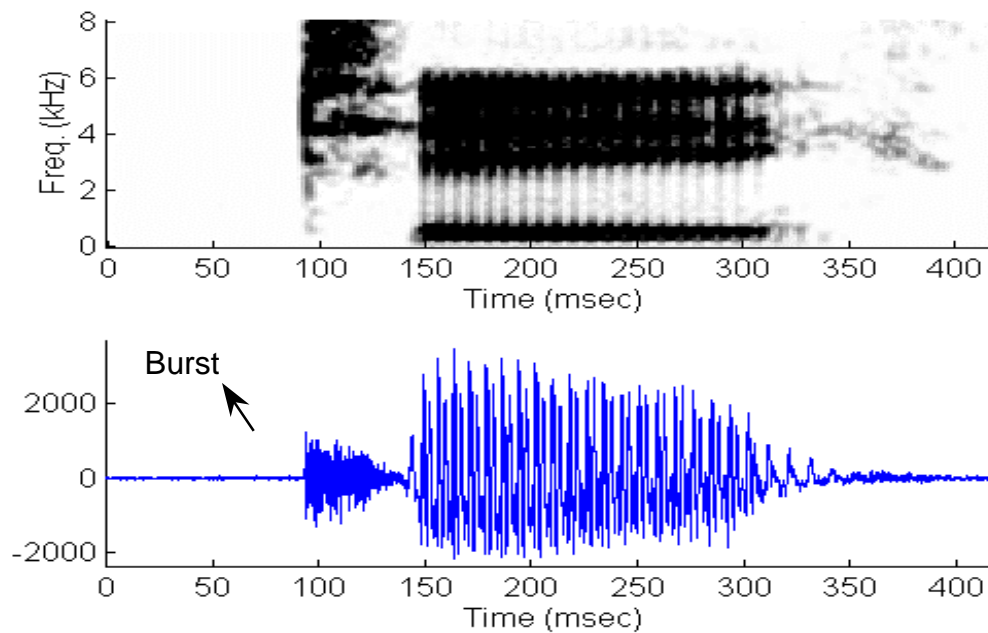
### **2.1.2 Stop consonants**

Consonants differ from the vowels in that they have higher energy in the higher frequency region than in the lower frequency regions as shown in Figure 2.10. General classification for the stop consonants is shown in Figure 2.3. As seen from the figure, stop consonants are one of the classes of consonants that are dealt in more detail in this thesis.

The essential feature of stop consonants is a momentary blockage of the vocal tract. The blockage is formed by an articulatory occlusion, which for English has one of the three sites: bilabial, alveolar, or velar. In English, the stop consonants are /p b t d k g/. Prevoalcalic stops have both a closure and a release phase. The articulatory blockage has a variable duration, usually between 50-100 ms and is subsequently released with a burst of air as the pressure impounded behind the obstruction escapes. Acoustically the closure phase is associated with a minimum of radiated energy. Because the vocal tract is obstructed, little or no acoustic energy is produced. Upon the release, a burst of energy is created as the impounded air escapes. Typically, the burst is no longer than 5-40 ms in duration and they are one of the shortest acoustic events that are commonly analyzed. Figure 2.11 shows the time waveform of a burst and its corresponding spectrogram. Stop releases are further classified into aspirated and unaspirated. Aspiration is a breathy noise generated as air passes through the partially closed vocal folds into the pharynx. The voiced stops in English are normally unaspirated. As mentioned earlier, the two main acoustic events of the



**Figure 2.10.** Time waveform of a consonant and its corresponding LPC spectrum showing high energy in the high frequency range.



**Figure 2.11.** Time waveform and its corresponding spectrogram showing the burst.

prevocalic stops are (a) stop gap or closure and (b) the release burst. The closure is the acoustic interval corresponding to the articulatory occlusion.

The transient that is produced at the release of the occlusion is known as a release burst and is no more than 40 ms in duration and is usually much shorter. It has been long recognized that the spectrum of a stop burst varies with the place of articulation. As Steven and Blumstein (1978) showed, labials tend to have low frequency dominance, alveolars are associated with high frequency dominance, and velars are characterized by mid-frequency burst.

## **2.2 Vowel Perception**

Vowel perception is relatively simple to understand since the positions of the first three formants relate directly to perception of different vowels. Four acoustic correlates to vowel perception are: (a) formant frequencies, (b) vowel duration, and (c) fundamental frequency. Contribution of each parameter to vowel perception is discussed below.

### **2.2.1 Formant frequencies**

Much of the experience with synthetic speech lends support to formant frequencies as a primary cue for vowel perception. Since the classic study of Peterson and Barney in 1952 on the distribution of the vowel formant frequencies on the F1-F2 plane, many studies were reported on the perception of vowels Strange (1989). A plot of the F1, F2 values for the ten vowels spoken by 60 speakers in an experiment conducted by Peterson and Barney

(1952) is shown in Figure 2.9. This study mainly focused to demonstrate strong relationship between the intended vowel and its formant frequency pattern. It was also demonstrated that there was considerable formant frequency variability from one speaker to the next for a particular vowel spoken and there was a substantial degree of overlap in the formant frequencies of adjacent vowels. However, the listening studies conducted showed that the vowels were highly identifiable in spite of the overlap in F1-F2 region of the adjacent vowels.

The above mentioned study by Peterson and Barney (1952), however had some limitations since there was no information regarding the dialects of either the speakers or listeners, individual results for men, women, and children and there was no way to identify individual tokens. But perhaps the greatest limitation was that the acoustic measurements were based on single time slice due to which there was no information available about the spectral change over time. Studies conducted by Ainsworth (1972), Bennett (1968), Jenkins et al. (1983) and Hillenbrand and Gayvert (1993) showed that dynamic properties like vowel duration and spectral change play an important role in vowel perception.

Hillenbrand *et al.* (1995) addressed the above mentioned limitation of Peterson and Barney studies. The average values of F1 and F2 obtained from the static measurements of the formants for the vowels were shown to occupy similar relative positions when compared to those obtained by Peterson and Barney (1952) with an exception of the vowel pair /ae/ and /ε/. Also, the importance of spectral change was demonstrated by Hillenbrand *et al.* (1995) from the fact that the phonemes /ae/ and /ε/ showed a considerable overlap in F1-F2 space and they could be separated by means of acoustic measurements only if spectral

change was considered. The studies showed that nearly all the vowels underwent formant frequency change. The formants moved in such a way so as to enhance the contrast between vowels with similar static positions in formant space. For example /ae/ and /ε/ showed a high degree of overlap when formants are sampled at steady state but exhibited distinct spectral change pattern. The influence of the spectral change patterns on the separability of the vowel categories can be seen from the results of a study reported by Hillenbrand *et al.* (1995). The study involved an identification algorithm to which the F1 and F2 values of vowels were the input. In the first case the only one sample of the vowel taken at steady state was given as an input to the algorithm and in the second case two samples of vowels taken at 20% and 80% of the vowel duration served as input. In the former, the accuracy of vowel identification algorithm was 76.1% and in the latter it was found out to be 90.3% which showed an increase in accuracy of around 14%. The studies by Hillenbrand (1995) and Hirahara (1992) also showed that other aspects of the vowels (e.g. F0, upper formants, bandwidths) presumably help listeners in vowel identification.

### **2.2.2 Vowel duration**

Although duration is neglected in the traditional F1-F2 chart, it is almost always available as a cue in the physical signal of speech. Among factors that influence the vowel duration are: tense-lax (long-short) feature of the vowel, vowel height, syllable stress and speaking rate. Experiments done by Hillenbrand *et al.* (1995) indicate that there was a consistent improvement in the identification accuracy when formant frequencies measurements were augmented with duration information. Although the duration is not

sufficient in itself to enable identification of any individual vowel, it does help the listener to distinguish spectrally similar vowels such as [ae] versus [e], as in “had” vs. “head” or “heed” vs. “hid” as shown by Sawusch (1997). In the same study conducted by Sawusch, it was shown that the effect of vowel duration was more pronounced when synthetic stimuli were used compared to natural stimuli.

### **2.2.3 Vowel fundamental frequency**

Vowels also vary in the fundamental frequency of phonation. Effects such as linguistic stress, speaker emotion and intonation often make the perceptual difference in the fundamental frequency less pronounced. There have been various studies investigating the effect of the fundamental frequency on vowel perception (Hirahara and Kato, 1992; Kewley-Port *et al.*, 1992; Hillenbrand *et al.*, 1995)

The Peterson and Barney (1952) experiment with vowel identification showed that vowels with the similar F1-F2 can be heard as different phonemes when uttered by different speakers having different fundamental frequency. These differences are to be expected from acoustic theory, in that the resonance frequency is determined by the length of the vocal tract which is different for various speakers. Also, they showed that the higher the fundamental frequency was, the higher the shift in the formant frequencies.

In a recent study by Hirahara and Kato (1992) the effects of pitch on vowel quality were examined using synthesized vowel stimuli whose formants as well as pitch were modified. It was found that pitch plays an important role in isolated vowel identification.

Apart from the formant frequencies, vowel duration and fundamental frequency (F0) which affect the perception of the vowels, formant bandwidth and amplitude also has a slight effect on the perception of the vowels. The concept of bandwidth can be associated with a formant since a formant, which is a peak in the spectrum corresponds to a resonance. In general any resonance can be described by two numbers, its resonance frequency and bandwidth. The primary perceptual effect of formant bandwidth is on the naturalness of the sound. Vowel that has unusually narrow formant bandwidth sounds artificial even though listeners can identify the vowel.

### **2.3 Consonant Perception**

Unlike vowel perception, consonant perception is an area of continuing controversy. While the acoustic cues leading to discrimination of manner of articulation are understood, the search continues for invariant cues to the feature of place of articulation. The acoustic similarity of stops with different place features has led to much discussion of the relative merits of formant transitions vs. burst frequency location.

In one of the first studies of acoustic cues for stop consonant recognition, Cooper *et al.* (1952) demonstrated that the frequency of the release burst, the onset frequency and the direction of change of the second formant transition were important cues to the recognition of the place of articulation. Steven and Blumstein (1978) described three distinct patterns corresponding to labials, alveolars, and velar stop consonants based solely on the shape of the spectrum sampled at 25 ms following the signal onset. They reported that for labial consonants, the number of peaks in the spectrum were fairly spread out or diffuse and the

amplitudes of the peaks either had more energy in the low frequency than the high frequencies (diffuse-falling pattern) or they were evenly distributed throughout the spectrum (diffuse-flat pattern). For alveolars, the spectrum of the release burst exhibited diffuse-rising pattern wherein the peaks were evenly distributed having larger amplitude in the high frequencies than the peaks in the lower frequency region. Finally for velar consonants, there was one prominent spectral peak, usually occurring in the mid-frequency range which dominated the entire spectrum (compact pattern). A high level of stop consonant identification was achieved (85% correct) based solely on the shape of the spectrum of the release burst (Stevens and Blumstein, 1978).

In other experiments, however, when listeners were presented signals with conflicting cues to the place of articulation, i.e., with onset spectrum specifying one place of articulation and formant transition specifying another, then the identification, in most instances, was based on the information provided by the formant transition (Dorman and Loizou, 1996). In an experiment conducted by Lahiri *et al.* (1984), the gross shape of spectrum failed to distinguish the labials and dental consonants in Malayam. This led to the realization that considering only the frequency spectrum of release burst did not yield any invariant acoustic cues across different languages.

Since the gross shape of the release burst spectrum failed to appropriately classify the stops across different languages, it was considered whether other properties inherent in the burst such as burst amplitude could be invoked. Lahiri *et al.* (1984) reported failure of burst amplitude to adequately classify the labial, dental and alveolar stops.

Kewley-Port (1983) explored the spectral change over time by focusing on the time-varying properties of stop consonants from the release burst into the vowel portions of CV syllables. One of the metric adopted was spectral tilt of the burst. The study done by Kewley-Port mainly focused on observing the absolute spectral tilt in different stop consonants in the release burst spectrum. But she did not take into account the relative changes in the spectral energy which occurred from burst to the formant transitions. It was then proposed by Lahiri *et al.* (1984) that the invariant properties for the labial and dental/alveolar place of articulation could be contained in relative changes in the distribution of energy at high and low frequencies when measured at the release burst and the beginning of formant transitions. Rather than looking at the absolute shape or the tilt of spectrum as Kewley-Port (1983) did, Lahiri *et al.* (1984) focused on the relative changes in the distribution of the energy from the burst release to the onset of voicing. Inspection of the three dimensional plots in which time was the third dimension revealed that the changes in distribution of energy from burst release to the onset of voicing were distinctively different for the labial and alveolar classes of stops.

The metric used in the Lahiri *et al.* study to distinguish between the stop classes was the ratio of the difference in energy of release burst at the onset of voicing at high and low frequencies. The LPC spectra of the burst and onset voicing were derived using a 10 ms window. The low-frequency marker was chosen at 1500 Hz and the high frequency marker was chosen at 3500 Hz. Slope for of onset voicing energy and the slope for burst energy at 1500 Hz and 3500 Hz was calculated. The ratio of former to the latter was computed. A positive ratio of  $< 0.5$  or negative ratio with numerator being negative, characterized dental

and alveolar stops, whereas a ratio of  $> 0.5$  or negative with denominator being negative indicated a labial stop.

Lahiri *et al.* (1984) reported that this metric both appropriately classified stops in a number of languages and functioned as a strong perceptual cue in the conflicting cue experiment. A study done by Dorman and Loizou (1996) verified that the relative spectral change was indeed effective in classifying the labials and alveolar stops in English. It was also verified by Dorman and Loizou (1996) that an attribute of the signal which allows accurate sorting of the signals by algorithm fails to exert a large influence on perception when other cues are present.

## **2.4 Spectral characteristics of vowels and consonants**

### **2.4.1 Vowels**

Vowels are associated with well-defined formant frequencies which have provided the dominant approach to acoustic characterization of these vowels. The Peterson and Barney's study helped us relate the vowel formant frequencies to vowel articulation. It was shown that F1 varies mostly as the tongue height and F2 varies mostly with the tongue advancement.

Modern speech synthesis often relies on formant frequency specifications of sounds to produce machine-generated speech. One advantage of description of vowels by formant frequencies is simplicity. In most cases, it is necessary to specify only the first three formants to obtain good quality vowels.

A full account of the acoustic cues for vowel perception would seem to require consideration of each of the following factors: formant frequencies, vowel duration, fundamental frequency and formant bandwidth. The shape of the vowel spectrum provides extra information regarding the perception of vowels. Spectral tilt in the spectrum of the vowel does not have a significant effect on the perception of the vowels. But a pronounced effect in vowel perception is observed if there is a shift in the relative position of spectral peaks. Hence the location of peaks and their movement due to addition of noise or any other reason may contribute to change in the perception of vowels. Vowel duration may help distinguish spectrally similar vowels whereas the fundamental frequency of the vowels may help distinguish the speaker. Formant bandwidth and amplitude can help perceive the naturalness of the spoken vowel. Yet another factor that may affect vowel identification is spectral contrast. Spectral contrast for a vowel is defined as the ratio of the maximum amplitude in the spectrum of the vowel to the minimum amplitude. Loizou and Poroy (2000) studied and reported the minimum spectral contrast needed for vowel identification by normal and cochlear implant listeners. The study revealed that although frequencies of the spectral peaks are considered to be the primary cues to vowel identity, the spectral contrast, i.e. the difference between the spectral peak and spectral valley, needs to be maintained to some extent for accurate vowel identification. Results of earlier study by Leek *et al.* (1987) showed that normal-hearing listeners required 1-2 dB peak-to-valley difference to identify four vowel-like harmonic complexes with relatively high (75%). Previous studies related to the spectral contrast, indicated that only a small spectral contrast is needed by normal hearing listeners for vowel identification. A larger spectral contrast is needed however, for hearing impaired listeners. Studies done by Leek *et al.* showed that

listeners with flat, moderate hearing loss required a 6-7 dB peak-to-valley difference for vowel identification. Studies done by Loizou and Poroy (2000) showed that Cochlear Implants (CI) listeners fitted with 6-channel Continuous Interleaved Strategy (CIS) processors, needed at least 4-6 dB of peak to trough ratio. The above studies suggest that spectral contrast is important for vowel identification by hearing impaired listeners. Because of the importance of spectral contrast and formant frequencies in vowel identification, we will study in this thesis the effect of noise on spectral contrast and formant frequencies.

#### **2.4.2 Consonants**

Consonants differ from vowels in that they had more energy in the high frequency region compared to the low frequency region. More emphasis was placed on the study of stop consonants. Stop consonants can be divided in three classes viz. labials, alveolars and velars, each having a distinct release burst spectrum shape.

For the stop consonants the peak in burst frequency revealed important information regarding the place of articulation. A peak in the spectrum of the burst at low frequencies was associated with the labials such as /b/ and /p/, whereas a peak at higher frequencies was found for alveolars such as /t/ and /d/. Velars such as /g/ and /k/ were found to have a peak in middle of the spectrum as shown by Steven and Blumstein (1978).

Studies by Steven and Blumstein (1978) showed that the spectral tilt of the release burst could also reveal information of the class of stop consonants. They reported that for labials, the burst spectrum had a diffuse-falling pattern. That is, the peaks in the spectrum were evenly spaced (diffuse). Also, the peaks in the lower frequency region had higher

energy than the peaks in the higher frequency region giving rise to a falling pattern. The burst spectrum for alveolars had a diffuse- rising pattern wherein the peaks were evenly spaced (diffuse) and/or the peaks at the higher frequencies had higher energy than those at lower frequencies. The burst spectrum of velars exhibited a compact spectrum which had high number of peaks were concentrated in the mid-frequency region than the low and high frequencies.

In brief, the major spectral characteristics of the stop consonants, important for identification, are the release burst frequencies, the shape of the burst spectrum and the formant transitions. In this thesis, we will study the effect of noise on the burst spectrum in terms of both spectral tilt and burst frequencies.

## **CHAPTER THREE**

### **ACOUSTIC ANALYSIS**

#### **3.1 Chapter outline**

This chapter deals with the acoustic analysis performed on the vowel and the consonant material. It provides explanation for the acoustic parameters used. This chapter is organized as follows. Section 3.2 deals with the speech material adopted for performing the acoustic analysis. Section 3.3 discusses the type of noises used and the motivation behind using those types. Section 3.4 explains the segmentation of the speech materials. Section 3.5 discusses the details regarding implementation of filterbank using critical band and logarithmic spacing. Section 3.6 and 3.7 discusses in detail the acoustic measurements for the vowels and stop consonants respectively. Details regarding the approach and implementation of measures are discussed.

#### **3.2 Speech material**

##### **3.2.1 Vowels**

Vowel material consisted of the vowels in the words: “heed, hid, hayed, head, had, hod, hud, hood, hoed, who’d, heard”. The stimuli were drawn from a large multi-talker vowel set used by Hillenbrand *et al.* (1995). This set is currently used extensively in the speech community for assessing vowel recognition of hearing-impaired listeners.

A total of 66 vowel tokens were used for acoustic analysis: 33 vowels produced by male speakers, and 33 vowels produced by female speakers. There were 6 tokens of each of 11 vowels, 3 produced by male speakers and 3 produced by female speakers. A total of 20 different male speakers and 23 female speakers produced the 66 vowel tokens. Each speaker produced only a subset of 11 vowels. The vowels used by Hillenbrand *et al.* were sampled at 16 kHz. Table 3.1 shows the mean values of the first two formant frequencies, F1 and F2 of all the vowels used in this study.

### 3.2.2 Consonants

Consonant material consisted of stop consonants. The stop consonants were in the form of VCV syllables. The words used were “aba, ada, aga, aka, apa, ata, ibi, idi, igi, iki, ipi, iti, ubu, udu, ugu, uku, upu, utu”. The stimuli were drawn from Shannon *et al* (1999).

A total of 36 consonant tokens were used for acoustic analysis: 18 consonants (6 stops x 3 vowel contexts) produced by a male speaker and 18 consonants produced by a female speaker. The consonants were sampled at 44.1 kHz.

### 3.3 Noise

Two types of noise were used: multi-talker babble (four talkers) and speech shaped noise. Unlike the white noise which has a flat frequency spectrum, the speech-shaped noise had frequency spectrum whose shape resembled the spectrum of speech.

Figure 3.1, shows the power spectral density of multi- talker babble and speech-shaped noise. The main motivation of using such type of colored noise was the fact that the real world acoustic noises are colored and not white.

		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 mean values of the first two formant frequencies (in Hz) of the male and female vowels used in this study.

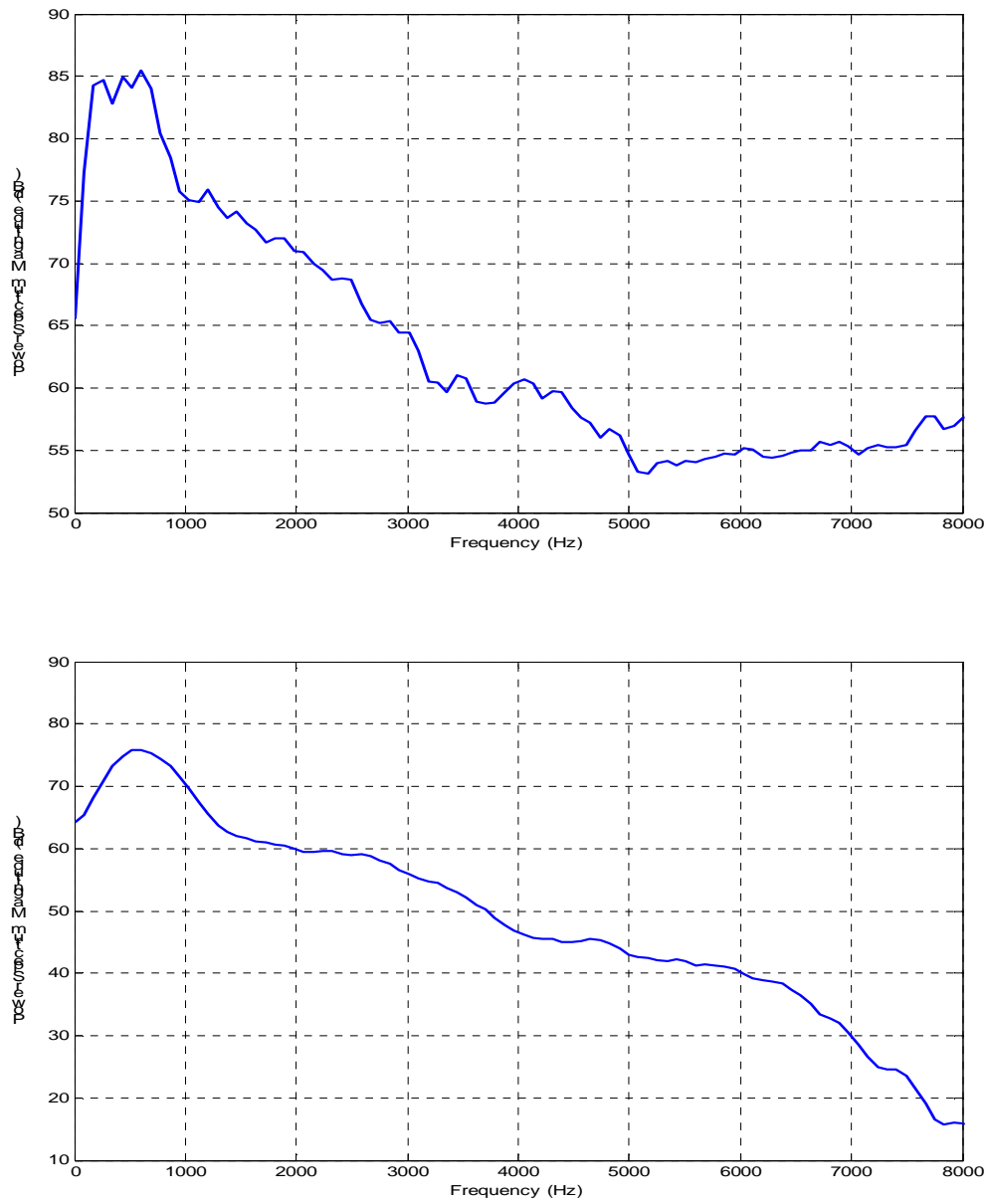
Both of the noises were added to the vowel and consonant materials at -5, 0, 5, 10 and 15 dB SNR.

### 3.3.1 Multi-talker babble

Sometimes referred to as cafeteria noise, the multi-talker babble was taken from the AUDiTEC CD (St. Louis) and was sampled at 22.05 kHz. Since this noise was to be added to the vowel material and the consonant materials mentioned in the previous section it was required to resample it. For the vowels, it was resampled to 16 kHz, and for consonants it was upsampled to 44.1 kHz.

### 3.3.2 Speech shaped noise

The speech-shaped noise was constructed by filtering white noise through a 60-tap FIR filter with a frequency response that matched the long-term spectrum of 11 male and 11 female vowels. As in case of multi-talker babble, necessary sampling was done.



**Figure 3.1.** (Top panel) PSD of multi-talker babble noise. (Bottom panel) PSD of speech shaped noise.

### 3.4 Segmentation

Prior to any analysis, the complete vowel data set was manually segmented to [h Vowel d]. The starting and ending times of the vocalic nuclei was measured by hand from high-resolution digital spectrograms. In order to avoid the effect of formant transitions due to /h/ and /d/, acoustic measurements were made starting from 20% of the vowel duration to 80% of the vowel duration. The vocalic segments were segmented into frames of 10 ms and acoustic analyses were done on these frames.

For the stop consonants, the release burst was analyzed over a 10 ms interval or the total burst duration, whichever was found smaller, starting from the burst. This was done in order to be consistent with the methodology followed in the studies by Steven and Blumenstein (1978), Lahiri *et al.* (1984) and Loizou and Dorman (1996).

### 3.5 Spectral analysis

#### 3.5.1 Filterbank approach

One technique for spectral analysis, popular due to the availability of real-time, simple, and inexpensive implementations, uses a filterbank or a set of bandpass filters, each analyzing a different range of frequencies of the input speech. Filterbank approach is more flexible than the DFT analysis since the bandwidths of the bandpass filters can be varied to follow the resolving power of the ear. Furthermore, the filter bank approach is particularly useful when a small set of spectral parameters describing the spectral distribution of energy are to be derived. Amplitude outputs from a bank of 24 bandpass filters typically provide a more compact and efficient spectral representation than a more detailed DFT.

Furthermore, the motivation to use filterbank for the spectral analysis of vowels in this thesis work stems from the fact that the results generated in this thesis could be used for enhancement of speech in cochlear implants. In cochlear implants, the electrodes are implanted in the inner ear at distances corresponding to the center frequencies of a 12-22 channel filterbank. Hence the results generated using this filterbank approach would be more appropriate for cochlear implants.

### **3.5.2 Critical band spacing**

In this thesis, a 21-channel filterbank was implemented using 6-th order Butterworth filters. The center frequencies of the filterbank were chosen according to critical-band spacing (Zwicker *et al.*, 1999). There are 21 critical bands in the 0-8 kHz range. The critical band spacing was chosen because the frequency response generated by the filter bank due to such spacing of the center frequencies of the band pass filters corresponds roughly to the tuning curves of auditory neurons. This fact was very important considering the possibility of application of the results generated in this thesis in the area of cochlear implants. The bandwidth of the filters in the lower frequency range is smaller and it progressively increases as the frequency range increases. Table 3.2 shows the frequencies and bandwidths associated with each critical band.

### **3.5.3 Logarithmic spacing**

In order to investigate the effect of number of channels on certain acoustic measurements like spectral contrast, it was required to perform the acoustic analysis using different number of channels, i.e. 4, 6, 8 and 12 channels.

Band	Lower edge frequencies (Hz)	Upper edge frequencies (Hz)	Center frequencies (Hz)	Bandwidth (Hz)
1	1	100	50	100
2	100	200	150	100
3	200	300	250	100
4	300	400	350	100
5	400	510	450	110
6	510	630	570	120
7	630	770	700	140
8	770	920	840	150
9	920	1080	1000	160
10	1080	1270	1170	190
11	1270	1480	1370	210
12	1480	1720	1600	240
13	1720	2000	1850	280
14	2000	2320	2150	320
15	2320	2700	2500	380
16	2700	3150	2900	450
17	3150	3700	3400	550
18	3700	4400	4000	700
19	4400	53000	4800	900
20	5300	6400	5800	1100
21	6400	7700	7000	1300

**Table 3.2.** Upper edge frequencies, lower edge frequencies, center frequencies, and bandwidths for 21-channel filterbank with critical band spacing.

Use of critical band spacing for small number of channels would be unsuitable since it would cover a very small region of the frequency spectrum. Hence, we used logarithmic spacing since it would approximate the human ear frequency response over the desired frequency range. Table 3.3 shows the center frequencies and bandwidths for 4, 6, 8, and 12 channel filterbank with logarithmic spacing.

	4-channels	6-channels	8-channels	12-channels
Lower edge frequencies in Hz	300, 620.8, 1248.5, 2658	300, 487.1, 791, 1284.5, 2085.8, 3387.1	300, 431.5, 620.8, 893, 1284.5, 1847.8, 2658, 3823.5	300, 382.3, 487.1, 620.8, 791, 1008, 1284.5, 1636.9, 2085.8, 2658, 3387.1, 4316.1
Upper edge frequencies in Hz	620.8, 1248.5, 2658, 5500	487.1, 791, 1284.5, 2085.8, 3387.1, 5500	431.5, 620.8, 893, 1284.5, 1847.8, 2658, 3823.5, 5500	382.3, 487.1, 620.8, 791, 1008, 1284.5, 1636.9, 2085.8, 2658, 3387.1, 4316.1, 5500
Center frequencies in Hz	460.4, 925.6, 1971.3, 4079	393.6, 639.1, 1037.8, 1685.2, 2736.5, 4443.5	365.8, 526.2, 756.9, 1088.7, 1566.1, 2252.9, 3240.7, 4661.7	3411, 4347, 5540, 7059, 8995, 1146, 1460.7, 1861.4, 2371.9, 3022.5, 3851.6, 4908.1
Bandwidth in Hz	320.8, 663.8, 1373.5, 2842	1871, 3039, 493.5, 801.3, 1301.2, 2112.9	131.5, 189.2, 272.2, 391.6, 563.2, 810.2, 1165.5, 1676.5	82.3, 104.9, 133.6, 170.3, 217, 276.5, 352.3, 449, 572.1, 729.1, 929.1, 1183.9

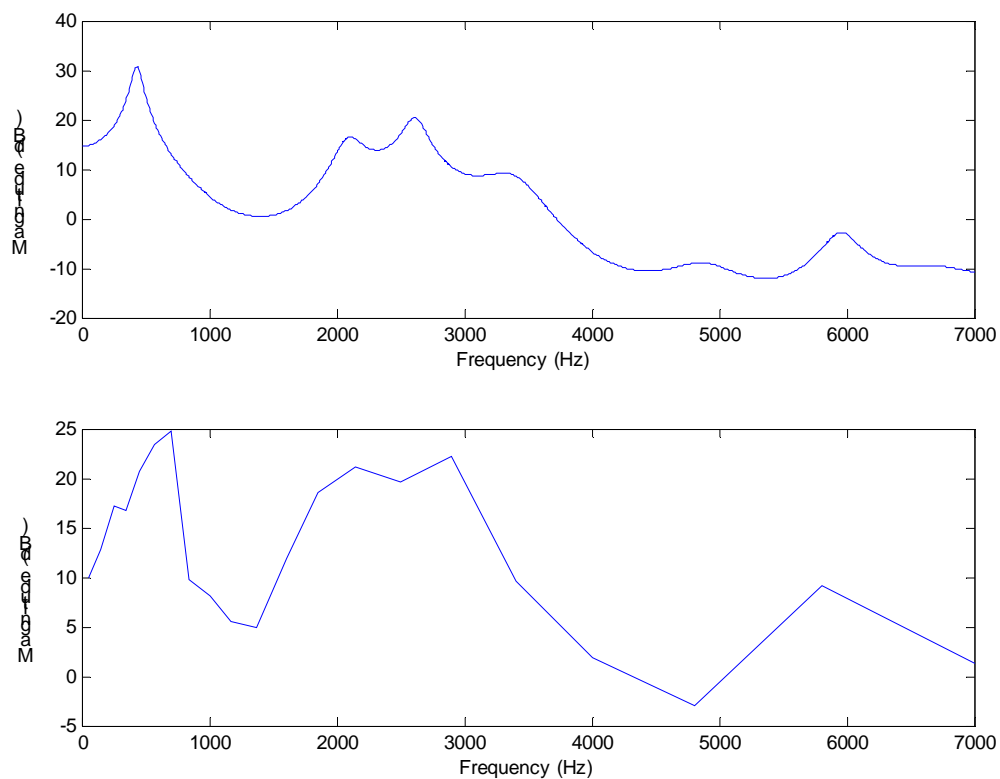
**Table 3.3.** Frequency spacing and bandwidths for 4, 6, 8, 12 channels with logarithmic spacing.

### 3.6 Acoustic measurements for vowels

#### 3.6.1 Spectral contrast measurements

The vocalic segment of the vowels (containing the 20% - 80% duration of the vowel duration) was first divided into frames of 10 ms. Each frame was first filtered through a 21-channel filterbank having center frequencies chosen in accordance to the critical band as discussed in section 3.5.

Estimates of the vowel spectra were then made by computing the root-mean-square



**Figure 3.2.** Plot of one sample frame (10 ms) of vowel /eh/. (Top panel) the 20-pole LPC spectrum. (Bottom panel) rms energy vs. frequency for a 21-channel filterbank.

(rms) energy of the 21 filterbank outputs. Measurements on spectral contrast were made every 10 ms, using the 21 filterbank values. Fig 3.2 shows as an example the rms energy plot for one 10 ms vowel frame.

Spectral contrast (in dB) was defined to be the difference between the spectral peak and the spectral valley, and was computed as follows:

$$SC_{dB} = 20 \log_{10} \frac{F_{\max}}{F_{\min}}$$

where  $F_{\max} = \max_{1 \leq i \leq 16} F_i$  is the spectral peak magnitude,  $F_{\min} = \min_{1 \leq i \leq 16} F_i$  is the spectral valley

amplitude, and  $F_i$  is the i-th rms filterbank value. The spectral peak and valleys were sought only within 0-3 kHz region, corresponding to critical bands 1-16. For most vowels the spectral valley lied between F1 and F2 (Hillenbrand *et al.*, 1995). Spectral contrast measurements were made for all vowels in quiet and all vowels in different noise conditions.

### 3.6.2 Spectral distance measurements

The vocalic segment of the vowels (containing the 20% - 80% duration of the vowel duration) was first divided into frames of 10 ms. Each frame was first filtered through a 21-channel filterbank having critical band spacing (Table 3.2). The spectral distance between the clean and noisy vowel spectra was then estimated for three different frequency bands (spanning the 0-8 kHz bandwidth) using the rms difference of the filterbank energies. The three bands considered were: 0-1 kHz (this is the band where F1 resides for most vowels), 1-2.7 kHz (this is the band where F2 resides for most vowels) and 2.7-8 kHz. Three spectral distance measurements were made, one of each band, every 10 ms:

$$\text{Low frequency band: } SB_1 = \sqrt{\frac{1}{9} \sum_{i=1}^9 (F_i^c - F_i^n)^2}$$

$$\text{Mid frequency band: } SB_2 = \sqrt{\frac{1}{6} \sum_{i=10}^{15} (F_i^c - F_i^n)^2}$$

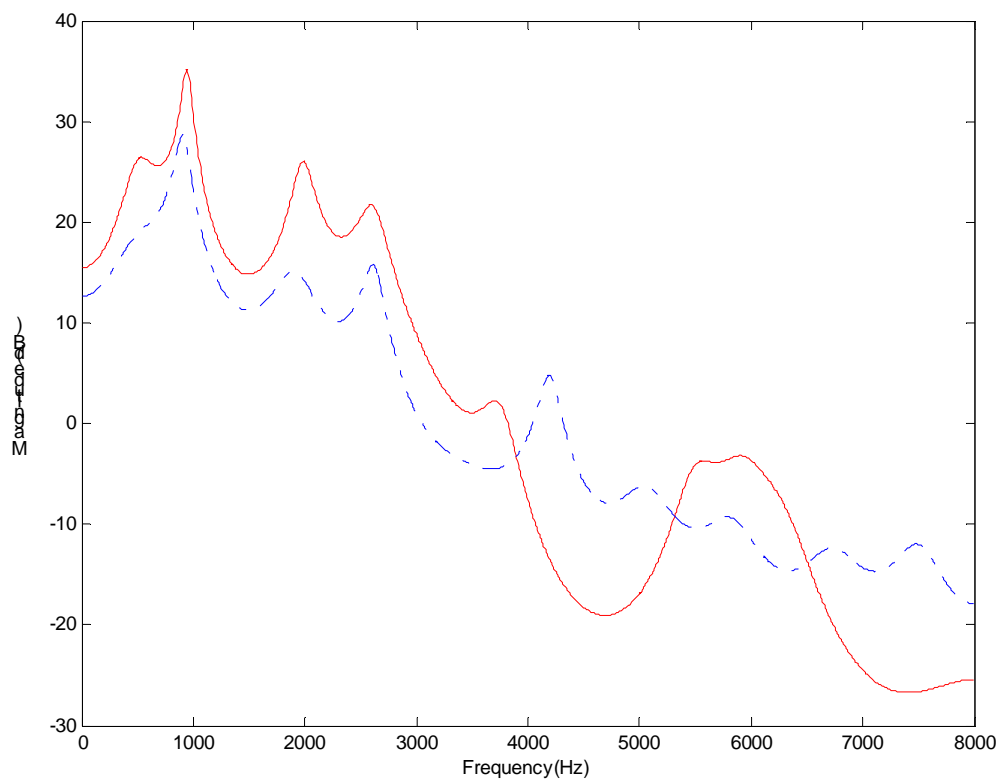
$$\text{High frequency band: } SB_3 = \sqrt{\frac{1}{6} \sum_{i=16}^{21} (F_i^c - F_i^n)^2}$$

where  $F_i^c$  denotes the  $i$ -th filterbank energy of the clean vowel, and  $F_i^n$  denotes the  $i$ -th filterbank energy of the noisy vowel.

### 3.6.3 Formant frequency measurements

Formant frequency measurements were made using a 22-pole LPC analysis over 20-ms hamming-windowed segments. The LPC spectrum was obtained using a 2048-point FFT, yielding a 7.8-Hz frequency resolution. The frequencies of the first seven spectral peaks were extracted from the LPC spectrum, every 20 ms. In order to get reliable F1 and F2 frequency estimates, formant frequencies were estimated manually rather than using a peak-picking algorithm. An interactive MATLAB program was used that allowed the user to select out of the first seven spectral peaks, the peaks corresponding to F1 and F2. The clean vowel spectrum was overlaid to the noisy vowel spectrum in order to get a rough estimate on the location of the F1/F2 frequencies of the noisy vowel spectra as shown in Figure 3.4. Knowledge of acoustic phonetics also played a role in the editing process, particularly the knowledge about the close proximity of F1 and F2 for vowels such as /a/ and /u/.

Although it is relatively easy to identify F1 and F2 in high SNR conditions, it is extremely difficult to identify F1 and F2 in extremely low SNR conditions. For that reason, F1 and F2 measurements were made only when it was felt that the selected peaks represented F1/F2 and not noise. For consistency purposes, several rules were adopted which classified each frame into 4 categories: (1) F1 not reliably detected, (2) F2 not reliably detected, (3) neither F1 nor F2 reliably detected and (4) F1 and F2 reliably detected. The percentage of frames that fell in each of the 4 categories was recorded for analysis.



**Figure 3.3.** LPC spectrum for a sample 10 ms frame of vowel /ae/. Plot shows 22-pole spectrum of clean and noisy speech at 0 dB. The spectrum with solid line represents the clean speech spectrum and the one with dashed line represents the noisy speech spectrum.

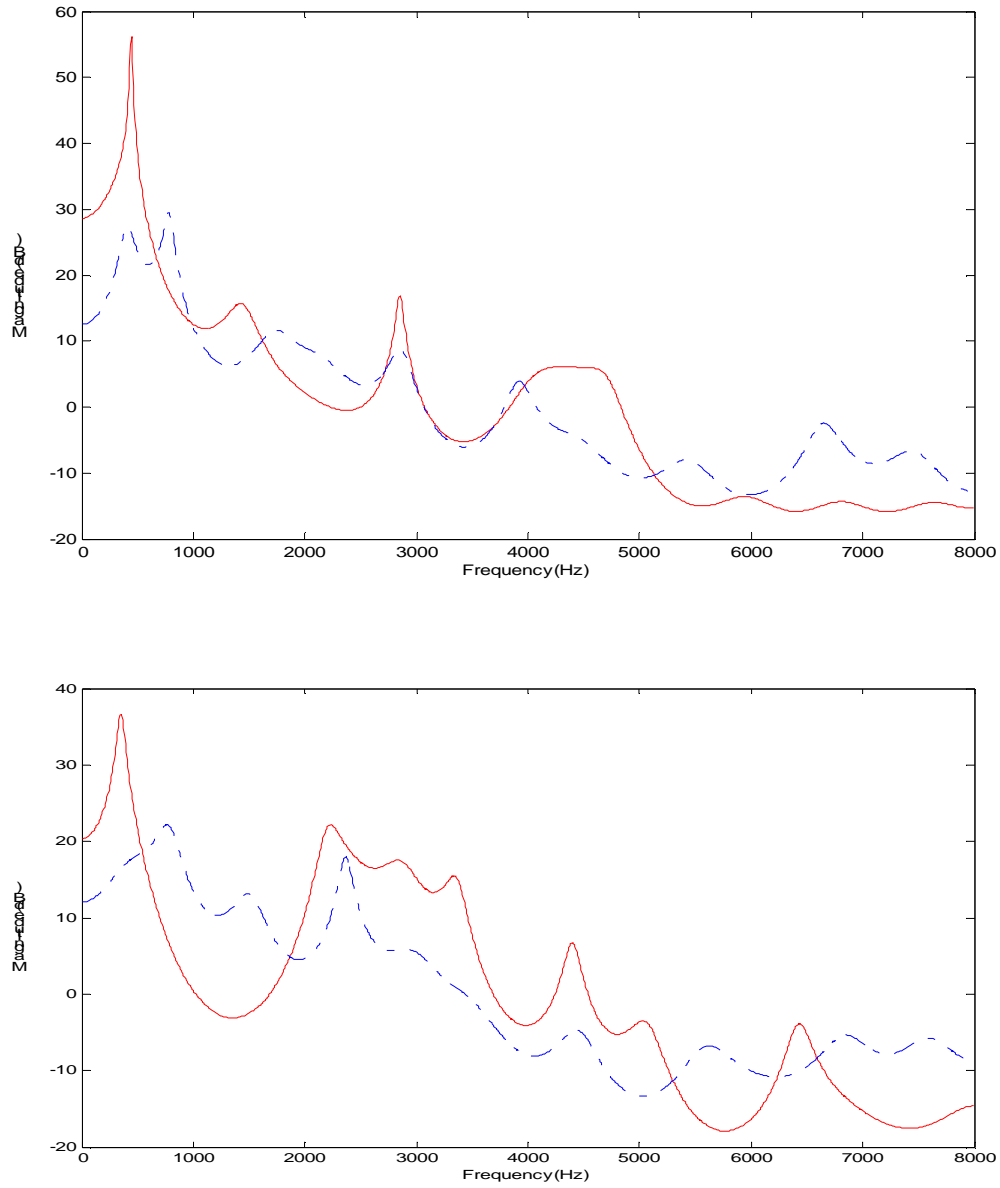
A frame was classified into category 1 (F1 not detected) whenever either of the following two conditions were satisfied (a) no peaks were detected in the proximity of F1 region, (2) two or more peaks were present in the proximity of the F1 region of the noisy vowel spectrum. A frame was classified into category 2 (F2 not detected) if either of the following three conditions were satisfied: (a) no peaks were detected in the proximity of F2 region, (b) two or more peaks were present in the proximity of the F2 region of the noisy vowel spectrum, (b) multiple peaks were present in the F1-F2 frequency range. A frame was classified into category 3 when it satisfied the conditions for both, category 1 and category 2. A frame was classified into category 4 whenever a single peak was found near the F1 and F2 regions. F1 and F2 measurements were made only for frames in this category.

Figure 3.4 to 3.7 show some examples for different category of frames. Figure 3.4 shows examples of two frames that can be classified as category 1 frames. Similarly, figures 3.5, 3.6, and 3.7 show some examples of frames that are classified as category 2, category 3, and category 4 frames.

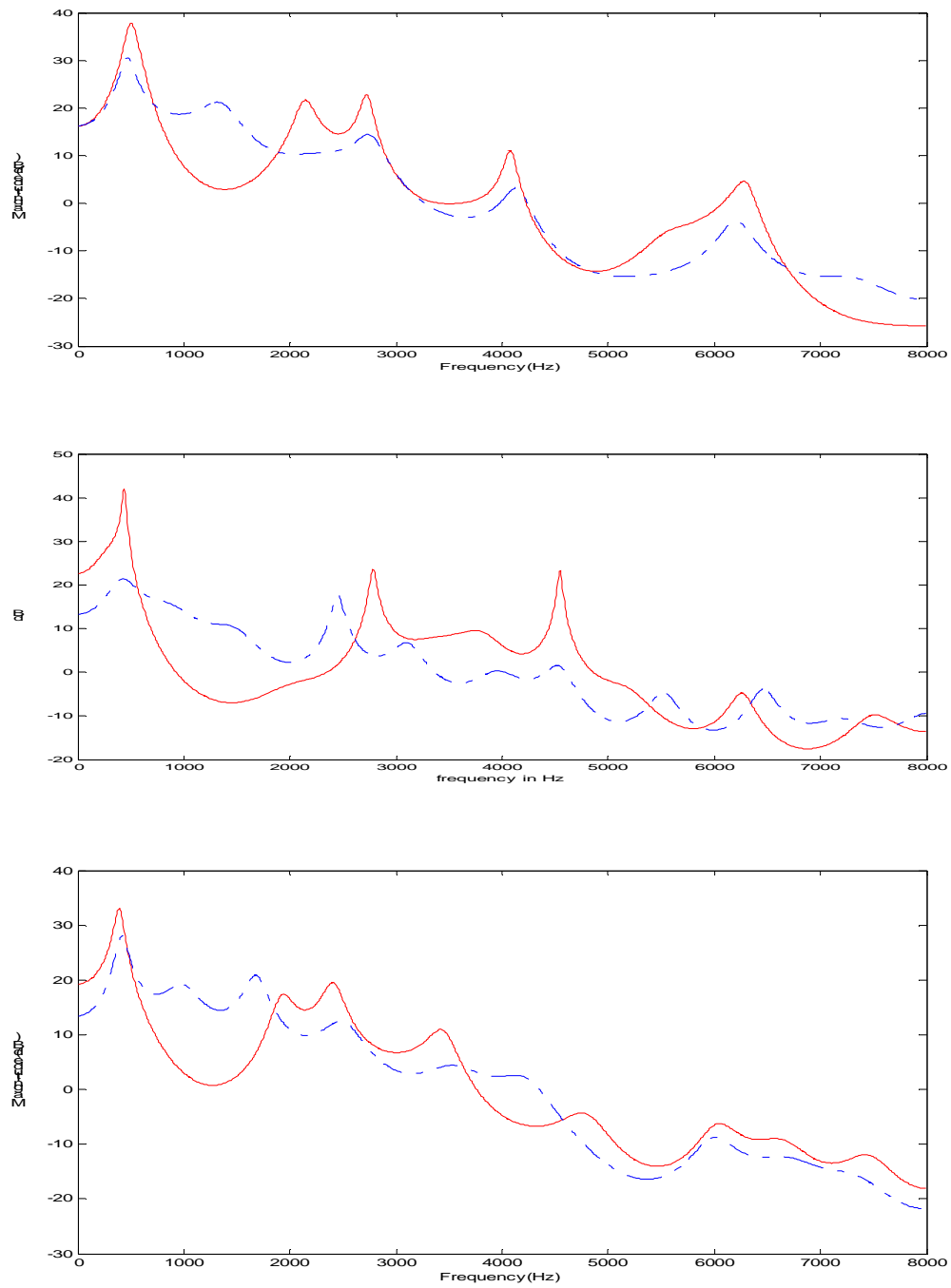
For frames in category 4 i.e. for the frames in which both F1 and F2 were reliably detected, difference between F1 and F2 ( $\Delta F1$  and  $\Delta F2$ ) was computed for quiet and noisy vowels.

### **3.7 Acoustic measurements for consonants**

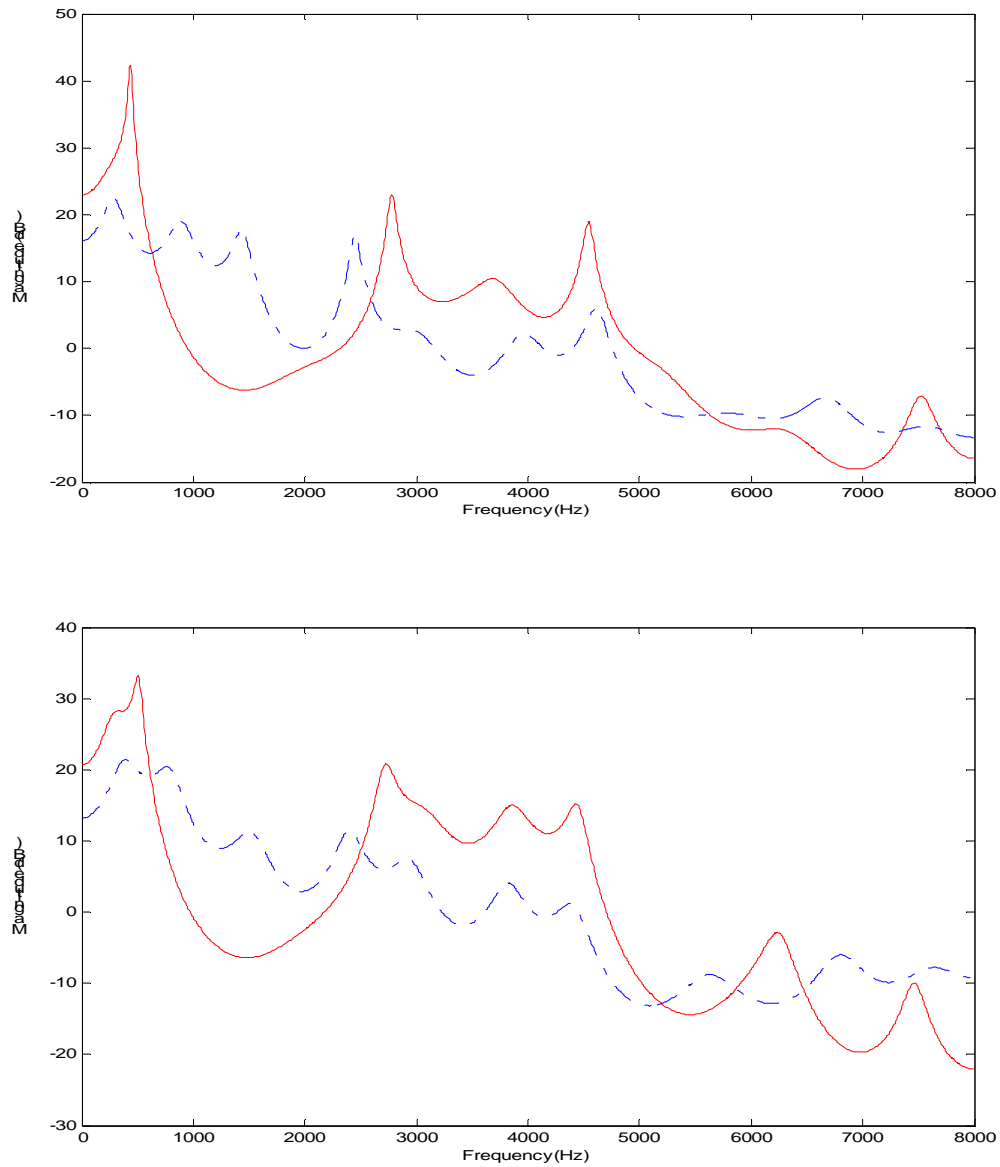
Prior to any analysis, the complete consonant data set was manually segmented to [Vowel Consonant Vowel]. The starting and ending times of the vocalic nuclei was measured by hand from high-resolution digital spectrograms. Only the first 10 ms of the release burst or the whole burst was considered.



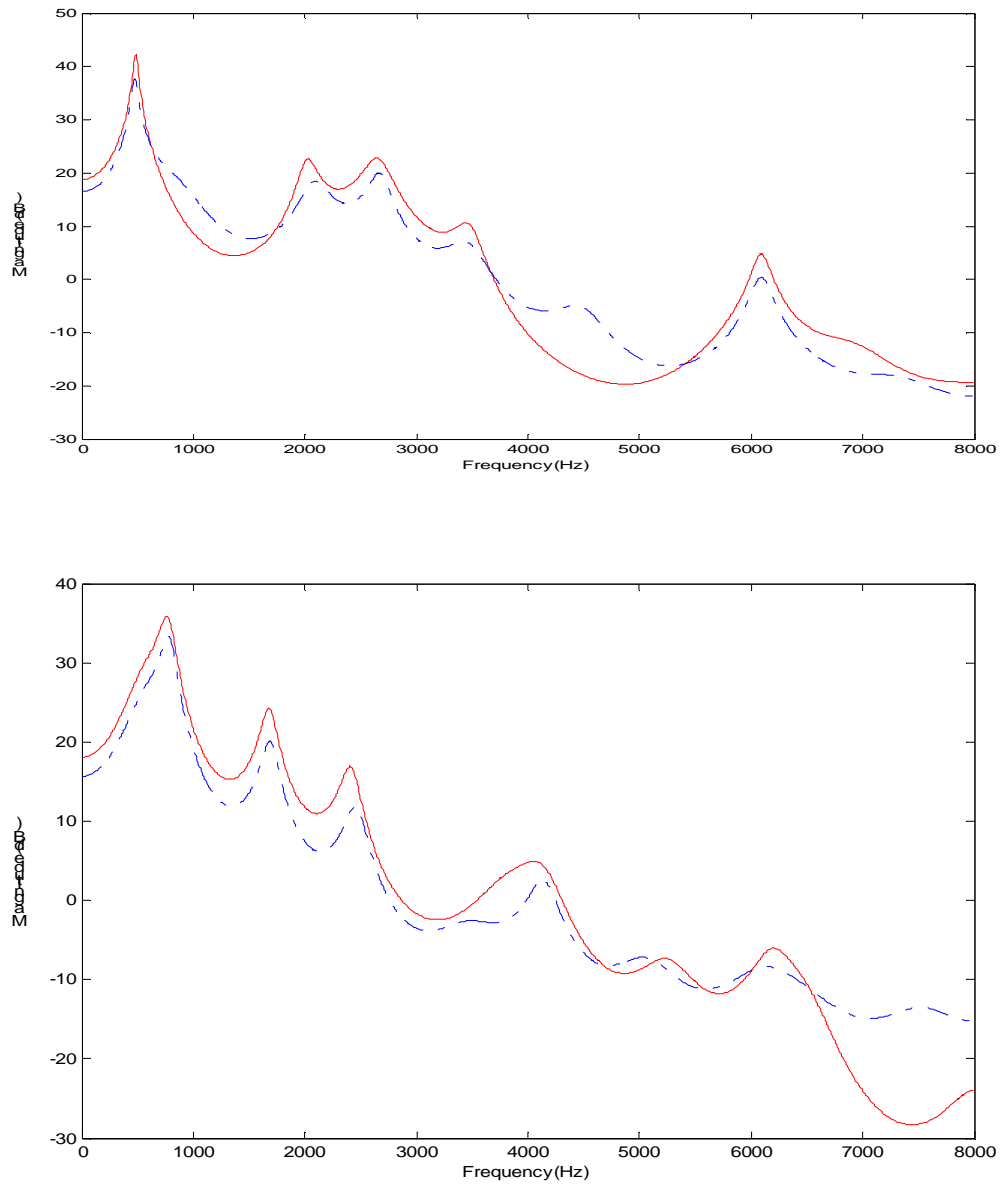
**Figure 3.4.** Examples of category 1 frame. (Top panel) Multiple peaks in the proximity of F1 region. (Bottom panel) No peaks detected in F1 proximity.



**Figure 3.5.** Examples of Category 2 frames. (Top panel) No peaks detected in the proximity of F2. (Middle panel) Multiple peaks in the proximity of F2. (Bottom panel) Multiple peaks between F1 and F2 region.



**Figure 3.6.** Examples of category 3 frames. Neither F1 or F2 can be detected in these frames.



**Figure 3.7.** Examples of category 4 frames. Both F1 and F2 can be reliably detected.

### **3.7.1 Measurement of the burst frequency**

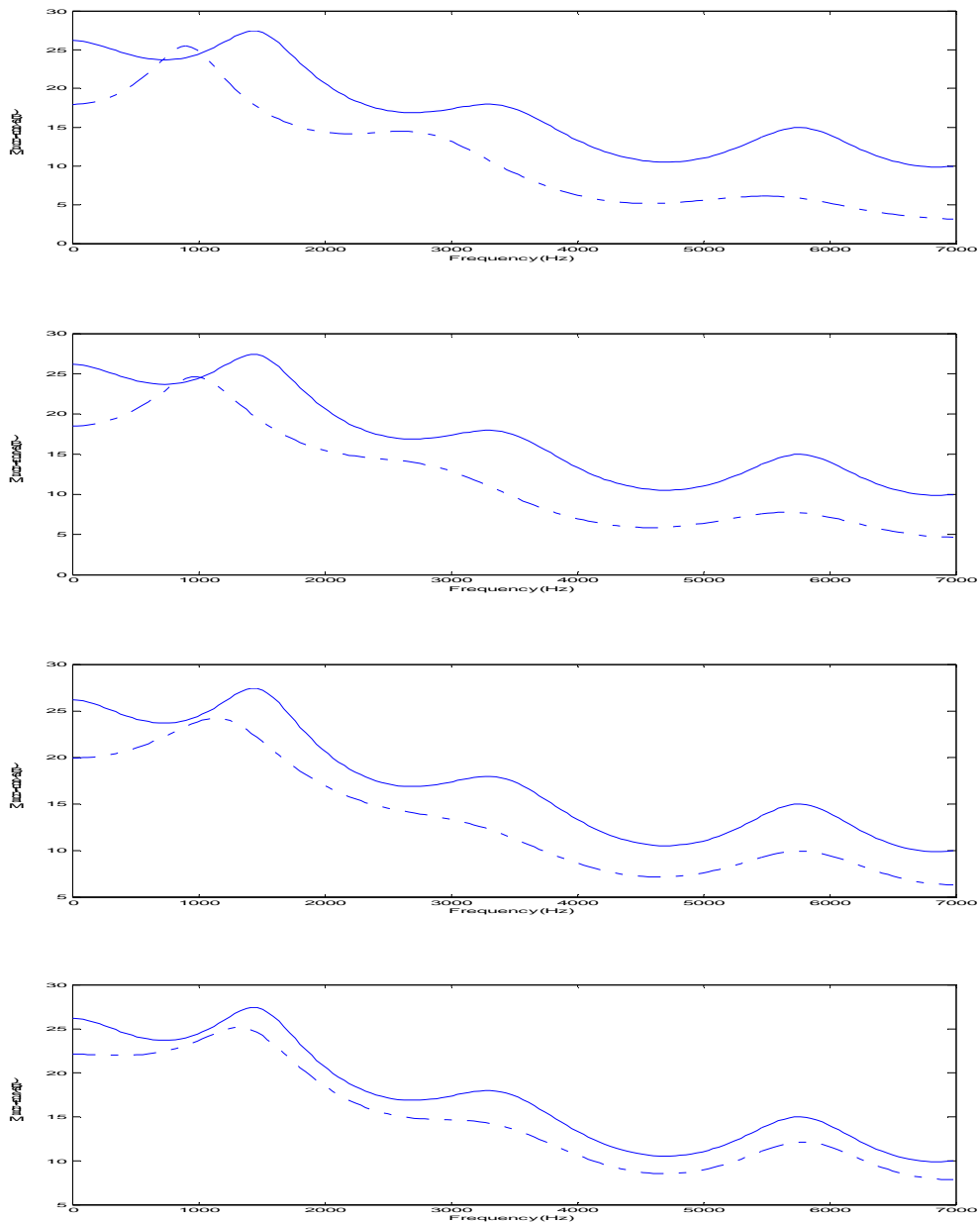
The burst frequency is the frequency at which the maximum amplitude in the frequency spectrum occurred. The burst frequency was estimated from the frequency spectrum of the release burst. Burst frequency measurements were made using a 20-pole LPC spectrum. A half-Hamming window of 20ms was used that is (only the latter half of the Hamming window was multiplied with the 10ms of the release burst samples). The LPC spectrum was obtained using a 512-point FFT.

The frequency of the maximum amplitude of spectral peak was extracted from LPC spectrum using a global peak picking algorithm. Once this acoustic analysis was done for both the clean and noise added consonants, the difference in the frequencies at which the peaks for the clean and noise added consonants occurred, was computed and plotted.

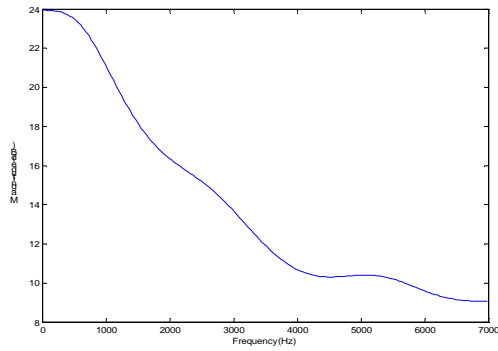
### **3.7.2 Measurement of tilt in the release burst spectrum**

As mentioned in the section discussing consonant perception in chapter 2, it was noted that the tilt of the release burst spectrum was very important in order to distinguish between the labials, velars and alveolars since each had a different shape of the burst spectrum (Steven and Blumstein, 1978). The shapes of the spectrum for labials, alveolars and velars are as shown in Figures 3.8. The motivation for this acoustic analysis is to find out whether the tilt of the release burst spectrum changes after adding noise. Figure 3.9 shows the spectrum of a stop consonant /ba/ after adding speech-shaped noise.

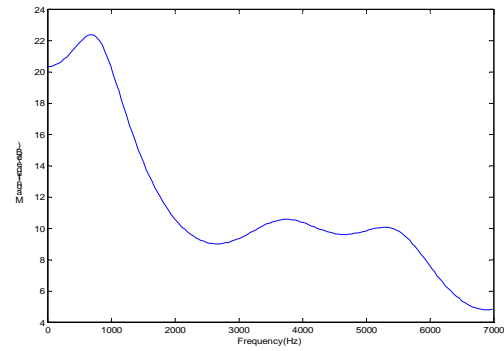
Burst frequency measurements were made using 4-pole LPC spectrum estimated using 512 point FFT. Such a low order of LPC was chosen because it was it was desired to get only the trend of the LPC spectrum over the desired frequency range and eliminate



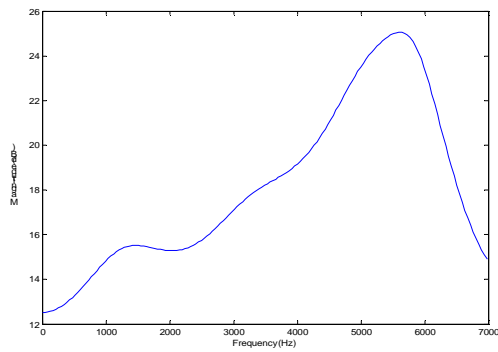
**Figure 3.8.** Examples of /ba/ 20-order LPC spectrum used to find the burst frequency. Solid lines show burst frequency spectrum for quiet /ba/ consonant. Dashed lines show the /ba/ burst spectrum under different noise conditions. (Top panel) -5 dB speech shaped noise condition. (Second panel from top) 0 dB speech shaped noise condition. (Third panel from top) 5 dB speech shaped noise condition. (Last panel) 10 dB speech shaped noise condition.



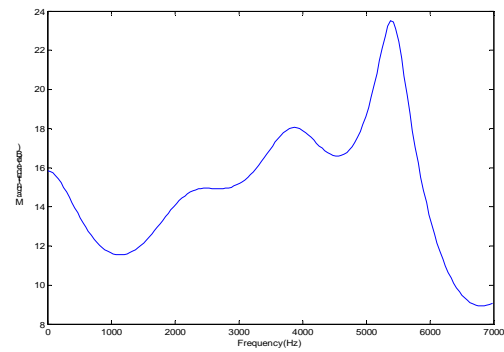
(a)



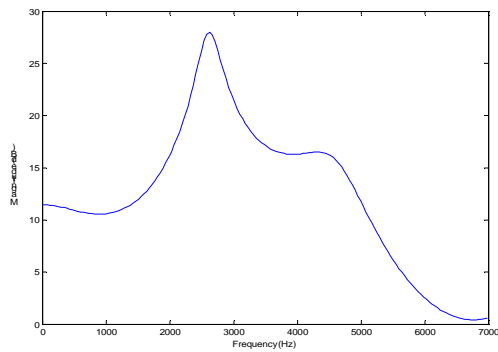
(b)



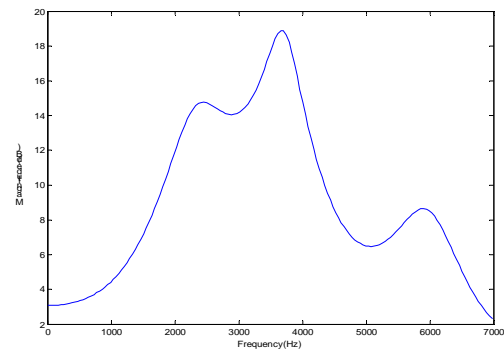
(c)



(d)



(e)



(f)

**Figure 3.9.** Plots of stop consonants derived using 20-pole LPC analysis are shown. Panels (a) and (b) show plots of labials /p/ and /b/ respectively. Panels (c) and (d) show plots of alveolars /d/ and /k/ respectively. Panels (e) and (f) show plots of velars /g/ and /k/ respectively.

unnecessary details in the spectrum.

To calculate the tilt in the spectrum, two points were used from on the LPC spectrum. It was decided to measure the tilt using the value of amplitudes at 1000 Hz and 5000 Hz in the spectrum.

### 3.7.3 Measurement of the spectral distance

In order to measure the change in the release burst spectrum once the noise was added to the consonants, spectral distance measure was used. A 21-channel filterbank was implemented using 6-th order Butterworth filters. The center frequencies of the filterbank were chosen according to critical-band spacing. After filtering the release burst through the 21-channel filterbank, the spectral distance between the clean and noisy burst spectra was estimated for three different frequency bands (spanning the 0-8 kHz bandwidth) using the root-mean-square difference of the filterbank energies. The three bands considered were: 0-1 kHz, 1-2.7 kHz and 2.7-8 kHz. The same three spectral bands were used for vowels. The spectral distance measurements were made, for each band, for a 10 ms frame of release burst:

$$\text{Low frequency band: } SB_1 = \sqrt{\frac{1}{9} \sum_{i=1}^9 (F_i^c - F_i^n)^2}$$

$$\text{Mid frequency band: } SB_2 = \sqrt{\frac{1}{6} \sum_{i=10}^{15} (F_i^c - F_i^n)^2}$$

$$\text{High frequency band: } SB_3 = \sqrt{\frac{1}{6} \sum_{i=16}^{21} (F_i^c - F_i^n)^2}$$

where  $F_i^c$  denotes the  $i$ -th filterbank energy of the clean consonant, and  $F_i^n$  denotes the  $i$ -th filterbank energy of the noisy consonant.

In the next chapter, we present the acoustic measurement of the vowels and consonants.

## **CHAPTER FOUR**

### **RESULTS**

#### **4.1 Chapter outline**

This chapter presents the results for the acoustic analysis performed on the vowels and the consonants as discussed in chapter 3. Each of the results generated are explained for various SNR and noise conditions.

This chapter is organized as follows. Sections 4.2, 4.3, and 4.4 discuss the results for the acoustic analysis of vowels while sections 4.5, 4.6, and 4.7 discuss the results of the stop consonants. For the vowels, spectral contrast results (section 4.2), spectral distance results (section 4.3) and formant frequency measurement (section 4.4) is discussed. For the stop consonants, spectral distance results (section 4.5), release burst frequency measurement (section 4.6) and measurement of tilt in the release burst spectrum (section 4.7) are explained.

#### **4.2 Spectral contrast for vowels.**

The spectral contrast was measured for the quiet vowels and noisy vowels after adding noise at -5, 0, 5, 10 and 15 dB SNR. Apart from measuring the spectral contrast using 21-channel filterbank, the measurements were made using 4, 6, 8, and 12 channels in order to investigate the effect of the number of channels on the spectral contrast. While critical band spacing was used for 21-channel filterbank, 4, 6, 8, and 12 channel

filterbank were implemented using logarithmic spacing. Both types of noise, speech-shaped and multi-talker babble, were used for calculation of spectral contrast.

Figures 4.1 and 4.2 shows the spectral contrast results of individual vowels using 21-channel filterbank for different SNRs for speech-shaped noise and multi-talker babble respectively.

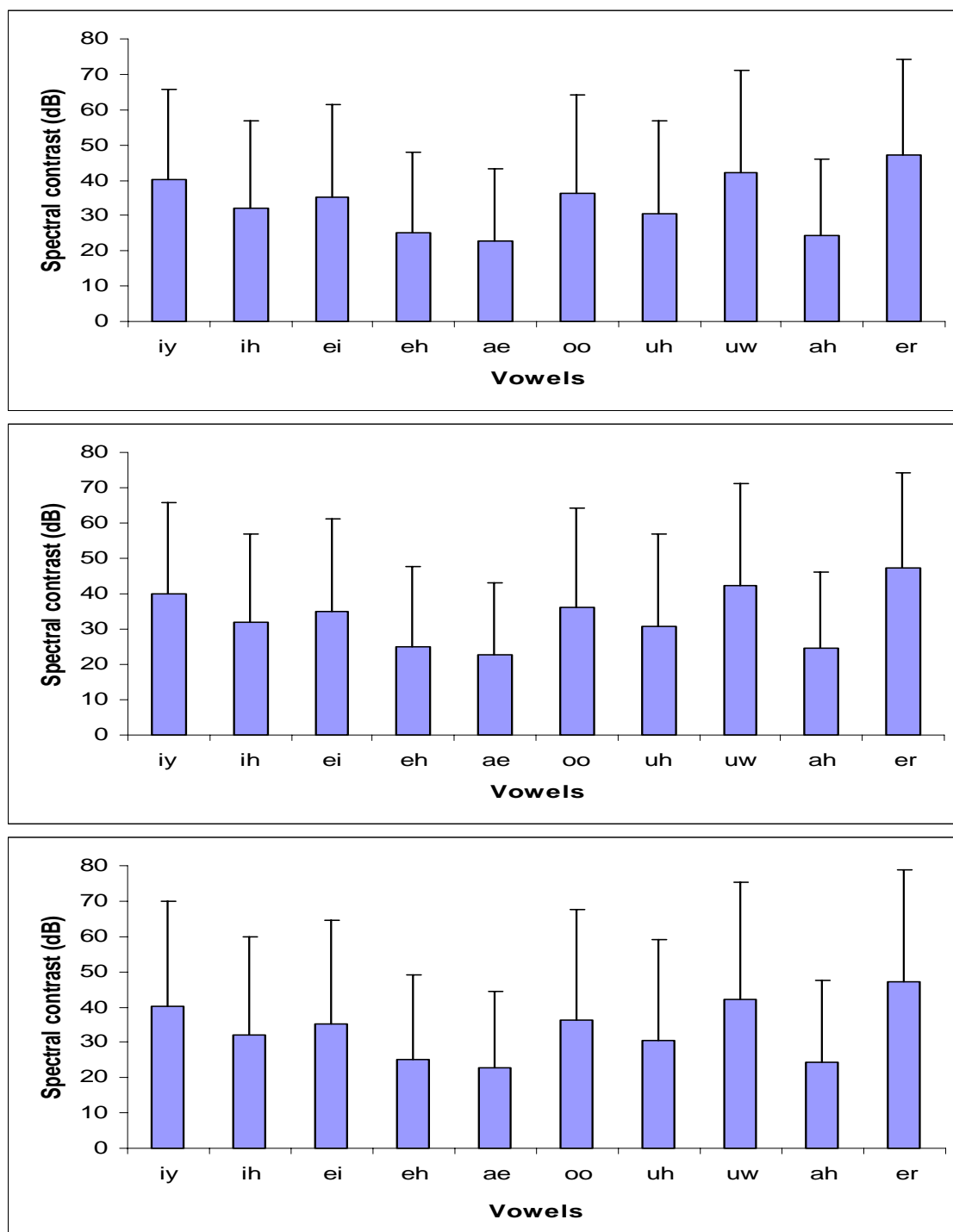
#### **4.2.1 Effect of number of channels on spectral contrast**

Figure 4.3 shows the effect of number of channels on the spectral contrast. As seen from the plot, the spectral contrast increases with the number of channels. Hence SNR remaining constant, the spectral contrast increases with the number of channels. There is a significant increase of about 10–12 dB in the spectral contrast when the number of channels are increased from 4 to 21 in case of speech-shaped noise and the corresponding increase in case of multi-talker babble is about 15-17 dB.

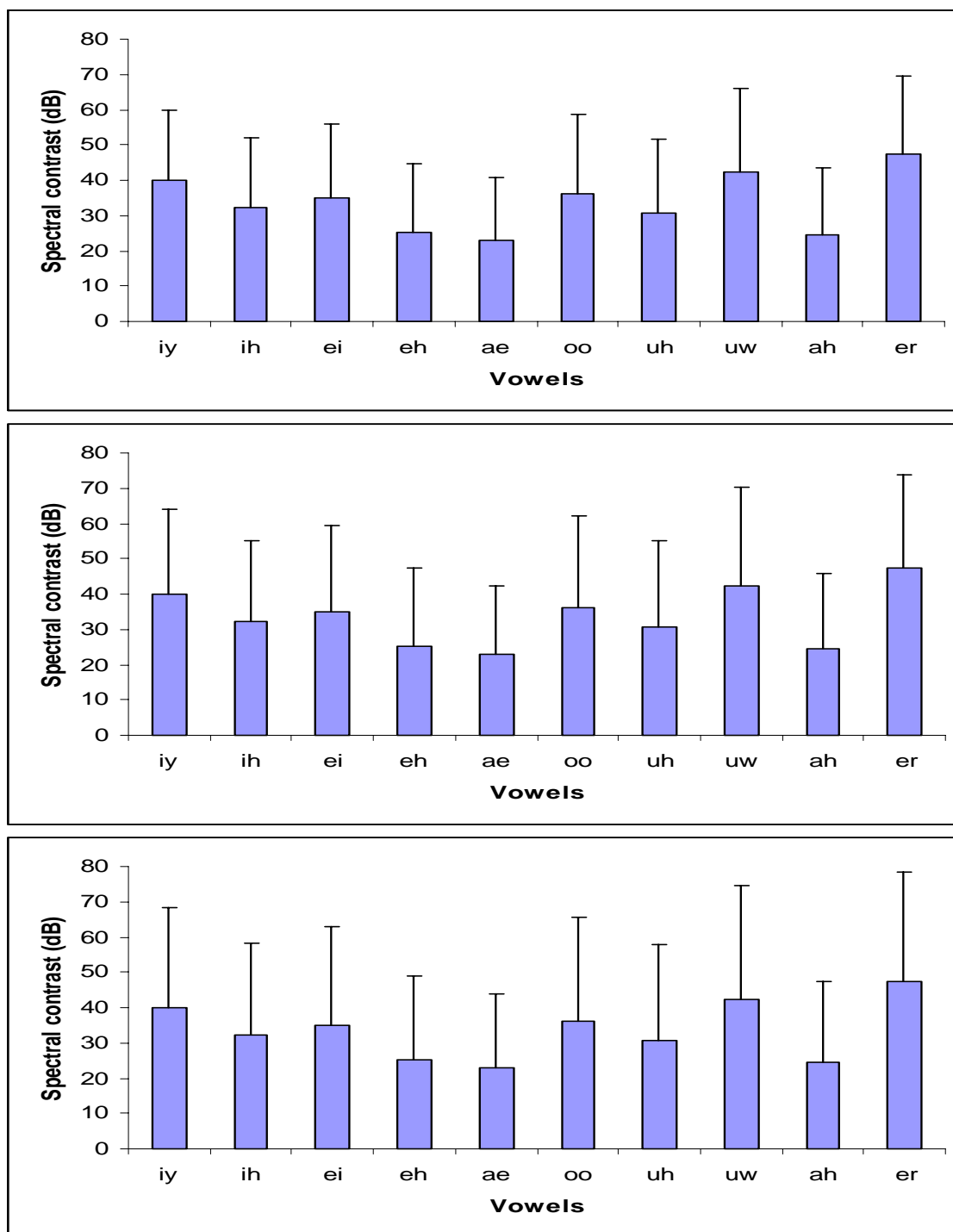
#### **4.2.2 Effect of type of noise on the spectral contrast**

Figure 4.4 shows the average spectral contrast measurements as a function of SNR for both types of noise. Spectral contrast decreased as the SNR decreased, and ranged from 18 dB at –5 dB SNR to 35 dB in quiet. This is in agreement with the notion that noise tends to flatten the spectrum of speech. There was a small difference of about 2 dB between the spectral contrast of vowels corrupted with multi-talker babble and vowels corrupted with speech-shaped noise.

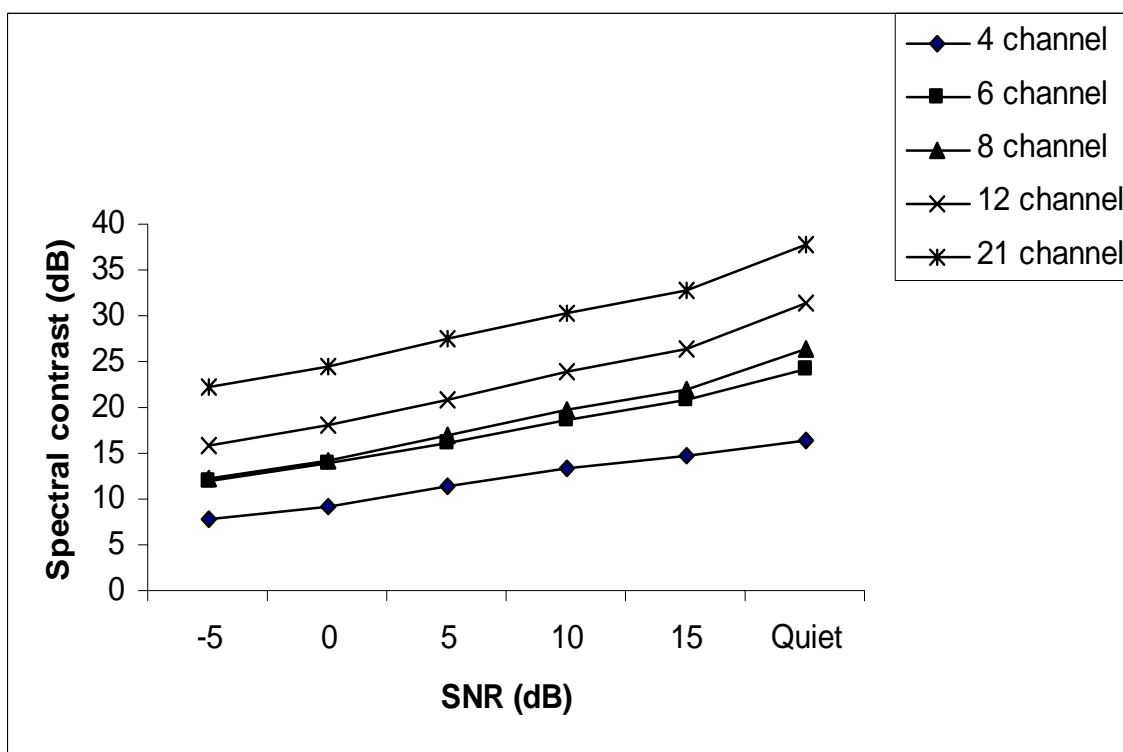
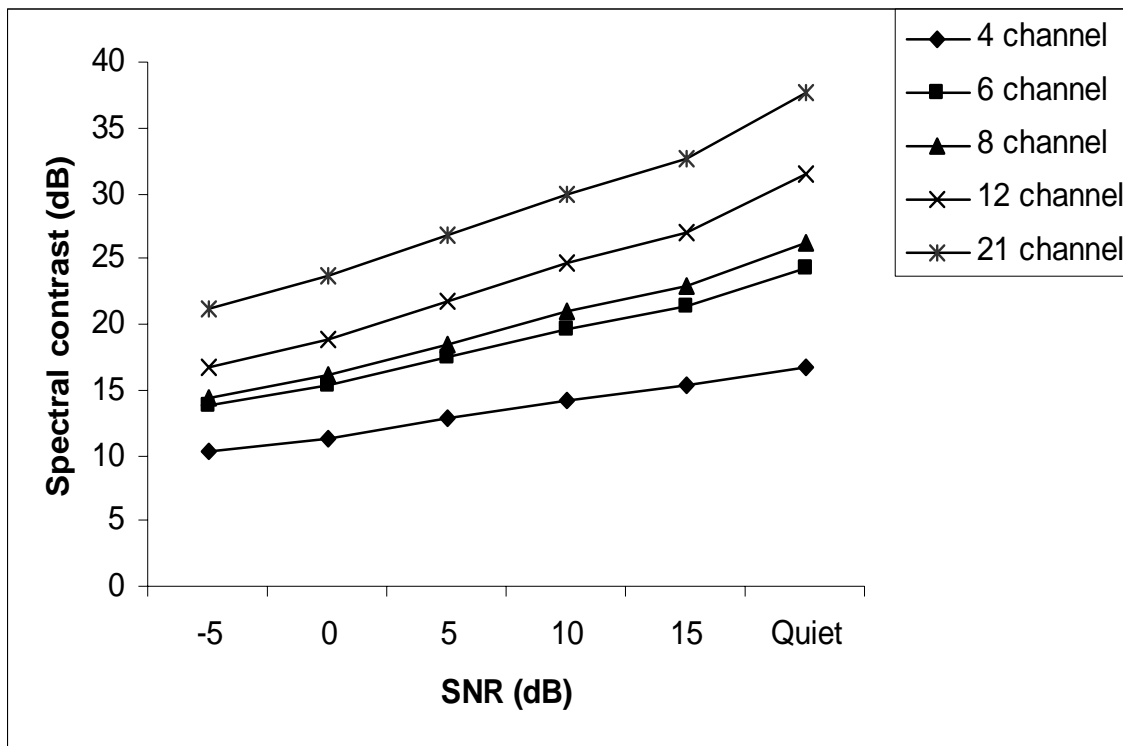
Hence in case of cochlear implants (CI), we can expect the CI listeners to perform well when there is higher number of channels used.



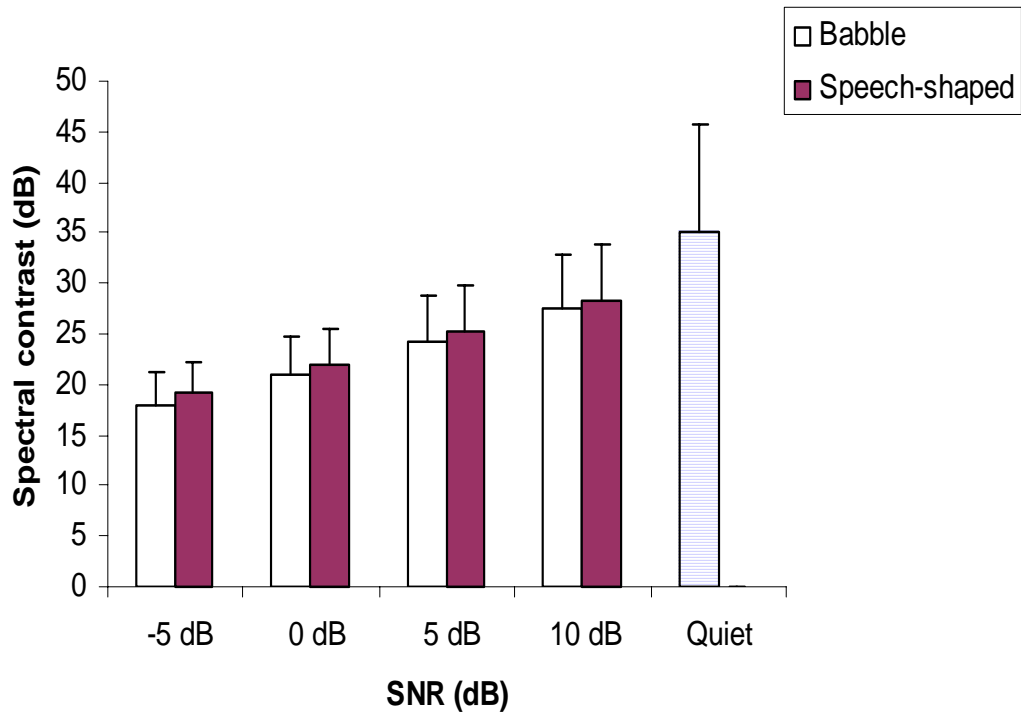
**Figure 4.1.** Spectral contrast for individual vowels corrupted by speech-shaped noise. (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise. The vowels used on the x-axis are: iy as in heed, ih as in hid, ei as in hayed, eh as in head, ae as in had, oo as in hood, uh as in hud, uw as in who'd, ah as in hod, and er as in heard.



**Figure 4.2.** Spectral contrast for individual vowels corrupted by multi-talker babble. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble. (Bottom panel) 10 dB multi-talker babble.



**Figure 4.3.** Effect of number of channels on spectral contrast. (Top panel) Results for speech-shaped noise. (Bottom panel) Results for multi-talker babble.



**Figure 4.4.** Comparison of spectral contrast for vowels corrupted by speech-shaped noise and multi-talker babble.

### **4.3 Spectral distance for vowels.**

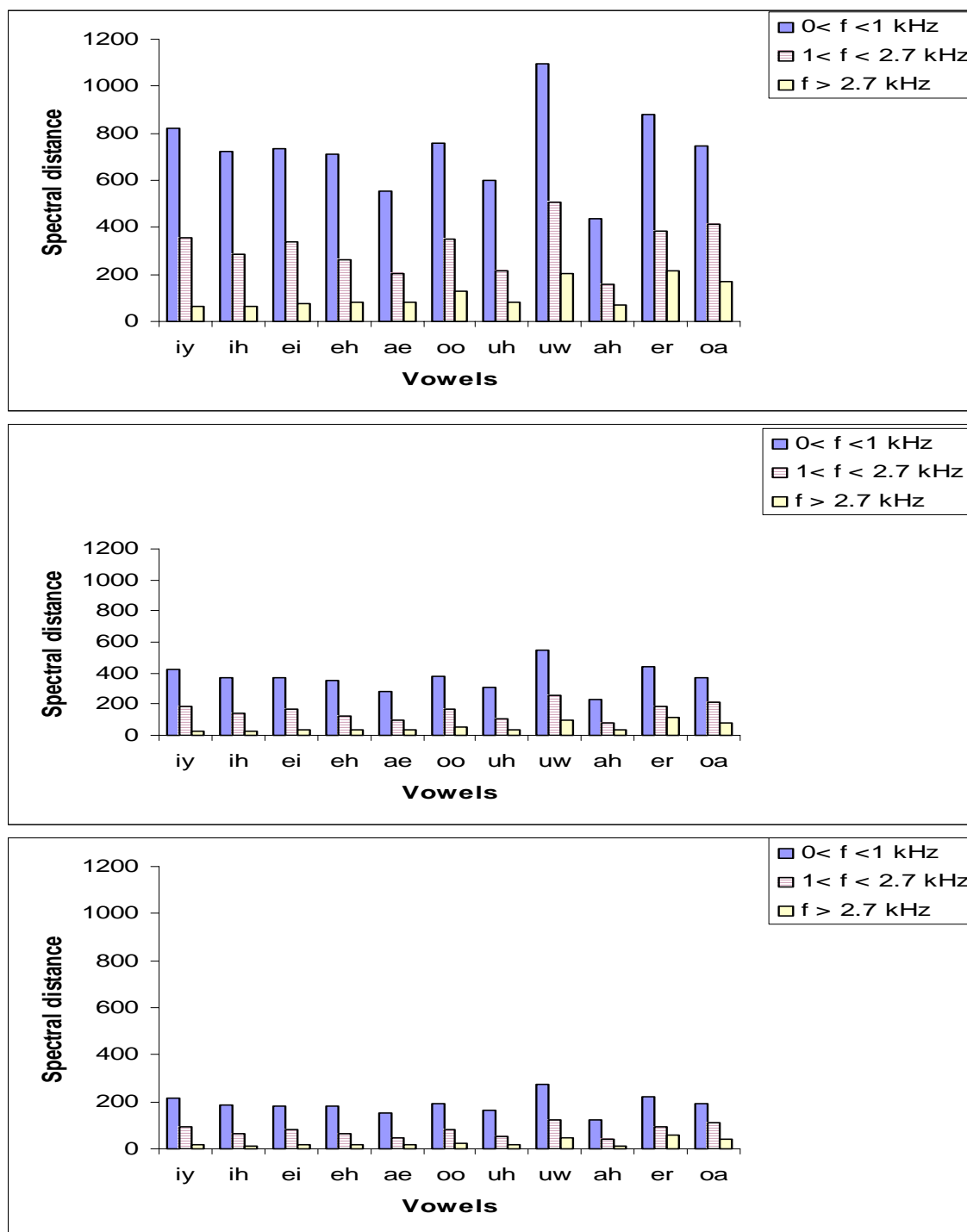
The spectral distance for the vowels was computed for both type of noises and for SNR conditions of -5, 0, 5, and 10 dB. Figures 4.5 and 4.6 show spectral distance results for individual vowels corrupted with speech-shaped noise and multi-talker babble respectively at different SNR.

#### **4.3.1 Effect of SNR on spectral distance**

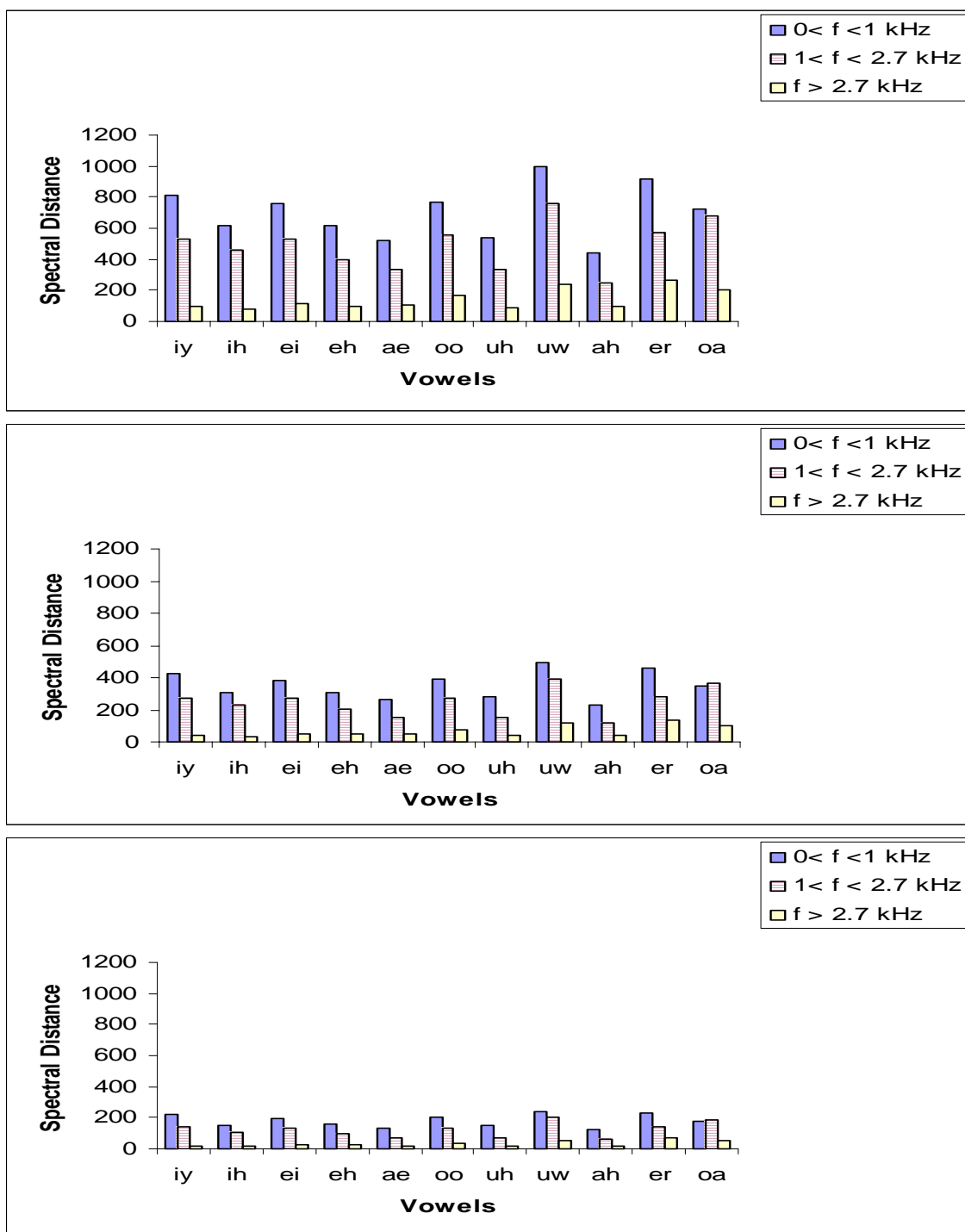
The average spectral distance measurements for the 3 frequency bands considered are shown in Figure 4.7. Overall, the spectral distance decreased as the SNR increased. The largest spectral difference between the noisy and clean vowel spectra occurred in the low-frequency band (0-1 kHz), followed by the middle-frequency band (1-2.7 kHz) and the high-frequency band (2.7-8 kHz). This pattern was consistent for both types of noise. The fact that the high-frequency band was affected the least by the noise was not surprising, since both multi-talker babble and speech-shaped noise affect primarily the low and mid frequency regions of the spectrum. This suggests that noise, at least the type considered in this study, affects the F1 region (0-1 kHz) of the spectrum to a larger degree compared to the F2 region.

#### **4.3.2 Effect of type of noise on spectral distance**

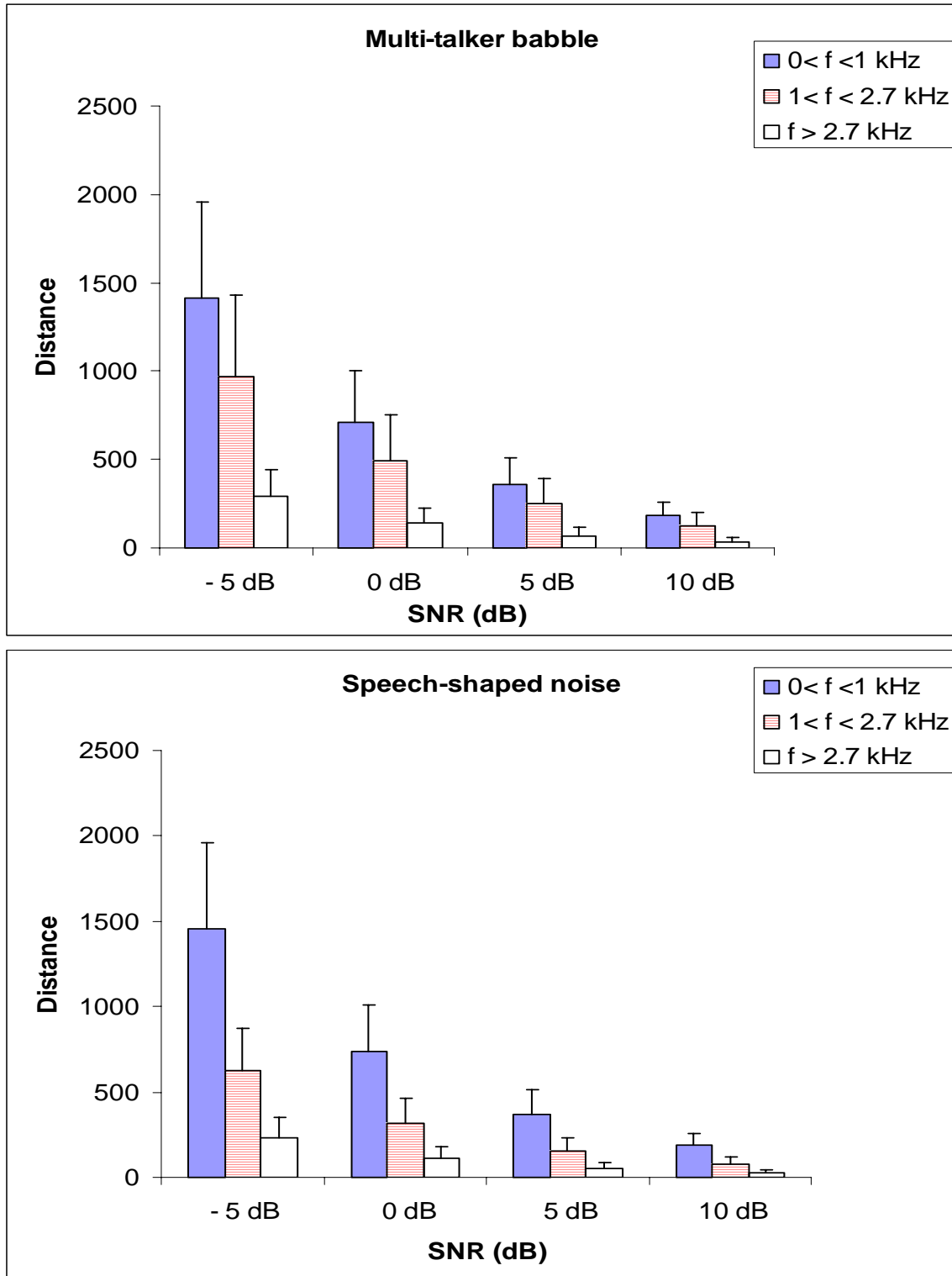
As shown in Figure 4.7, multi-talker babble affects the F2 region of the spectrum to a larger extent compared to speech-shaped noise.



**Figure 4.5.** Spectral distance results for individual vowels corrupted by speech-shaped noise. (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise.



**Figure 4.6.** Spectral distance results for individual vowels corrupted by multi-talker babble. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble. (Bottom panel) 10 dB multi-talker babble.



**Figure 4.7.** Average spectral distance for vowels. (Top panel) Spectral distance results for speech-shaped noise. (Bottom panel) Spectral distance results for multi-talker babble.

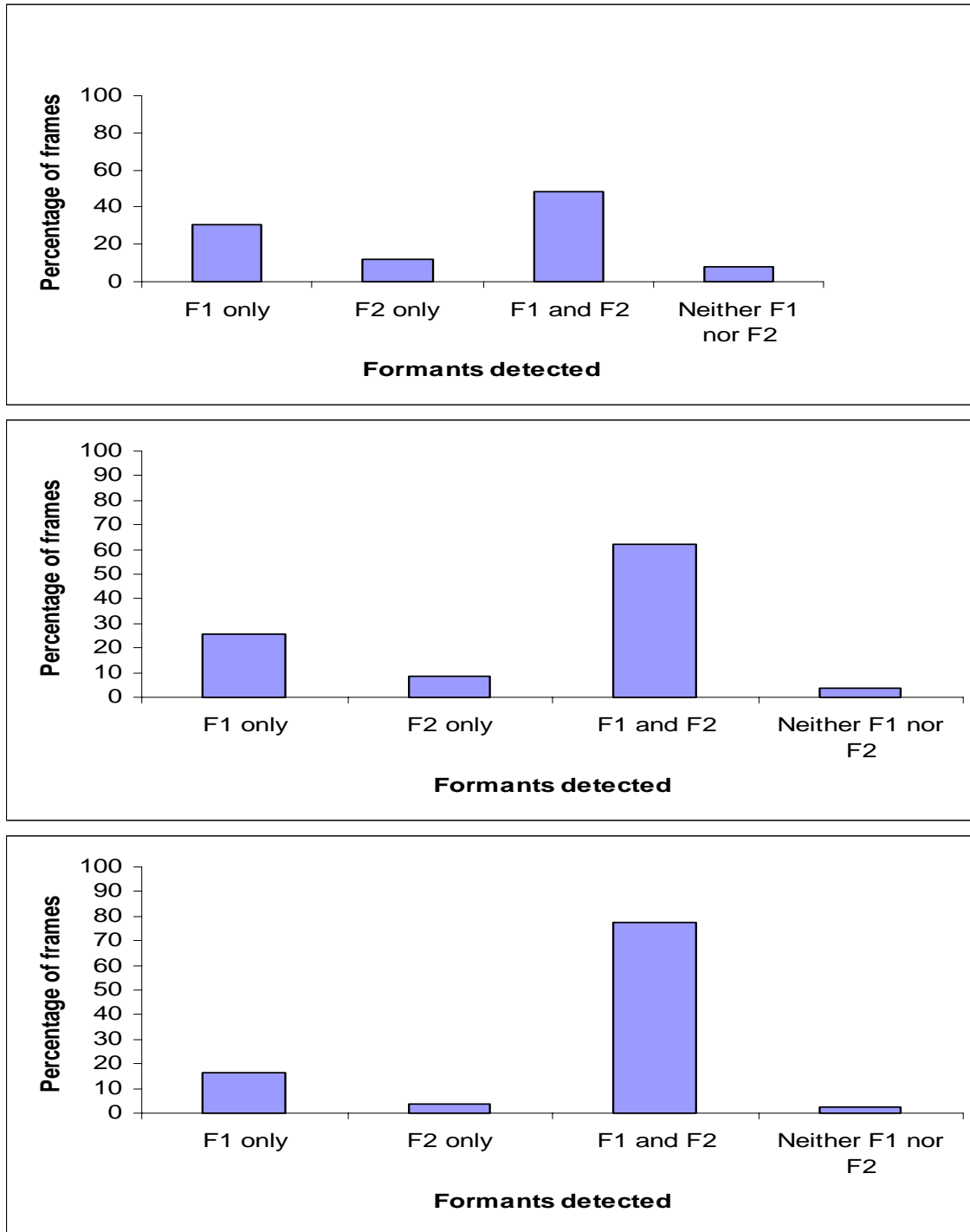
#### 4.4 Formant frequency measurement for vowels.

The formant measurements for the vowel material were computed for both types of noises and for -5, 0, 5, and 10 dB SNR conditions. A total of 539 frames (240 frames for men and 299 frames for women) were used in the analysis.

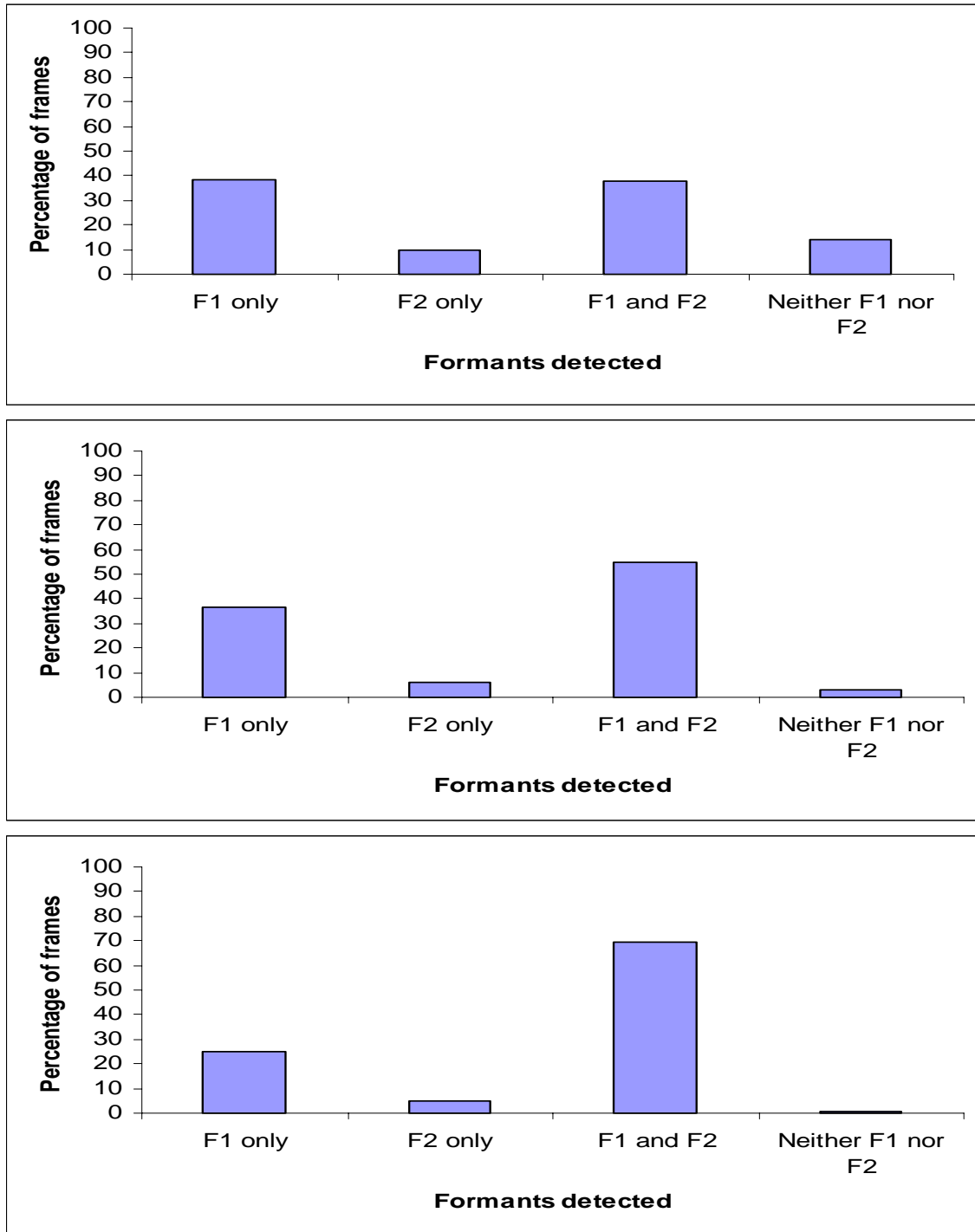
Figures 4.8 and 4.9 show results for formant frequency measurements done for speech-shaped and multi-talker babble noise respectively.

Figure 4.10 shows the percentage of frames (out of a total of 539 frames) for which F1 and/or F2 were reliably detected for vowels embedded in speech-shaped noise (top panel) and for vowels embedded in multi-talker babble (bottom panel) at various SNRs. For the low SNR conditions, the first formant (F1) was reliably detected more often compared to the second formant (F2). For instance, in multi-talker babble at -5 dB S/N, F1 was detected 60% of the time while F2 was detected only 30% of the time [these values were obtained by summing the percentages of “F1 only” and “F1&F2” in Figure 4.8 in case of speech-shaped noise and Figure 4.9 in case of multi-talker babble]. In speech-shaped noise at -5 dB S/N, F1 was detected 64% of time, while F2 was detected 48% of the time. The fact that F2 was detected less often than F1 is consistent with the spectral distance measurements which showed that the multi-talker babble affects the F2 region of the spectrum to a larger degree than the speech-shaped noise does.

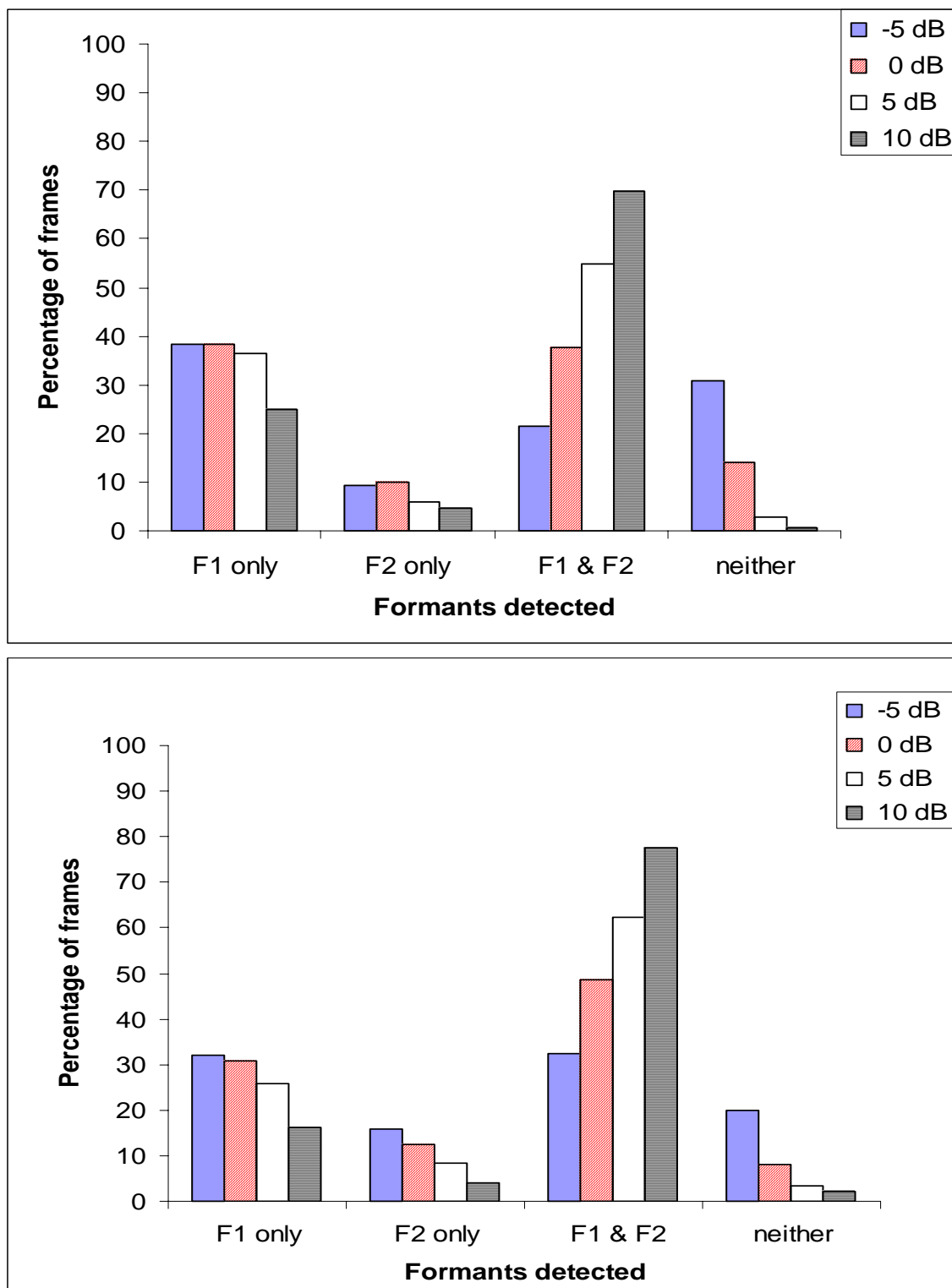
Both F1 and F2 formants were reliably detected more than 50% of the time only for SNR values of 5 dB and higher. In multi-talker babble at 5 dB S/N, F1 and F2 formants were detected 54.6% of the time, while in speech-shaped noise the two formants were detected 62% of the time.



**Figure 4.8.** Formants detected for speech-shaped noise. (Top panel) 0 dB. (Middle panel) 5 dB. (Bottom panel) 10 dB.



**Figure 4.9.** Formants detected for multi-talker babble. (Top panel) 0 dB. (Middle panel) 5 dB. (Bottom panel) 10 dB.



**Figure 4.10.** Formants detected for different SNR. (Top panel) Results for speech-shaped noise. (Bottom panel) Results for multi-talker babble.

#### 4.5 Formant frequency deviation of vowels.

Formant frequency deviation can have a profound effect on vowel perception as discussed in chapter 3. Hence it was desired to know the range of deviation in formant frequencies under different noises and SNR conditions. To perform this analysis, the data computed in the previous section, for the formant frequency measurement was used.

For the frames in which both formants were reliably detected, additional analysis was performed to determine how close the formant frequencies estimated in noise were compared to the estimated formant frequencies in quiet. The absolute difference between the formant frequencies was computed as follows:

$$\Delta F_1 = | F_1^q - F_1^n |$$

$$\Delta F_2 = | F_2^q - F_2^n |$$

where the superscript  $q$  indicates the corresponding formant frequency in quiet, and the superscript  $n$  indicates the formant frequency estimated in noise. The results are tabulated in Table 4.1 for SNR=5 and SNR=10 dB.

As can be seen from Table 4.1, the differences in formant frequencies ( $\Delta F$ s) are extremely small. These  $\Delta F$ s are close to the difference limens (DLs) of formant frequencies. A difference limen is the maximum frequency deviation that can be tolerated between the first and the second formant so as to maintain the same perception of the vowel. Studies by Flanagan (1955) showed that difference limens were estimated to be in the order of 1-2% of the formant frequencies.

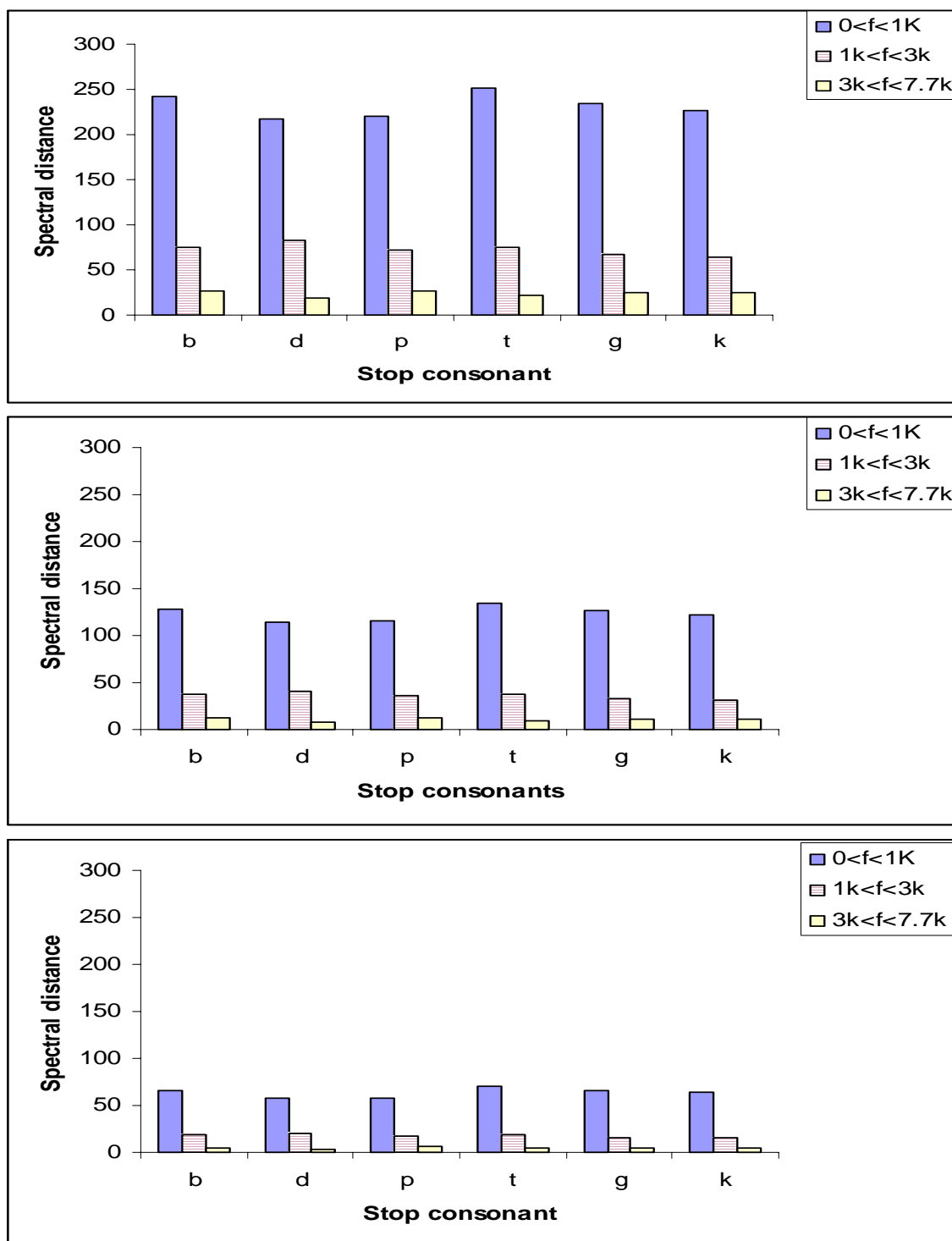
Noise	$\Delta F_1$ (Hz)	$\Delta F_2$ (Hz)
Multi-talker: SNR=5 dB	19.3	27.6
Multi-talker: SNR=10 dB	12.4	20.0
Speech-shaped: SNR=5 dB	15.8	22.8
Speech-shaped: SNR=10 dB	11.1	16.9

**Table 4.1.** Mean  $\Delta F$  values between formant frequencies of corrupted and clean vowels.

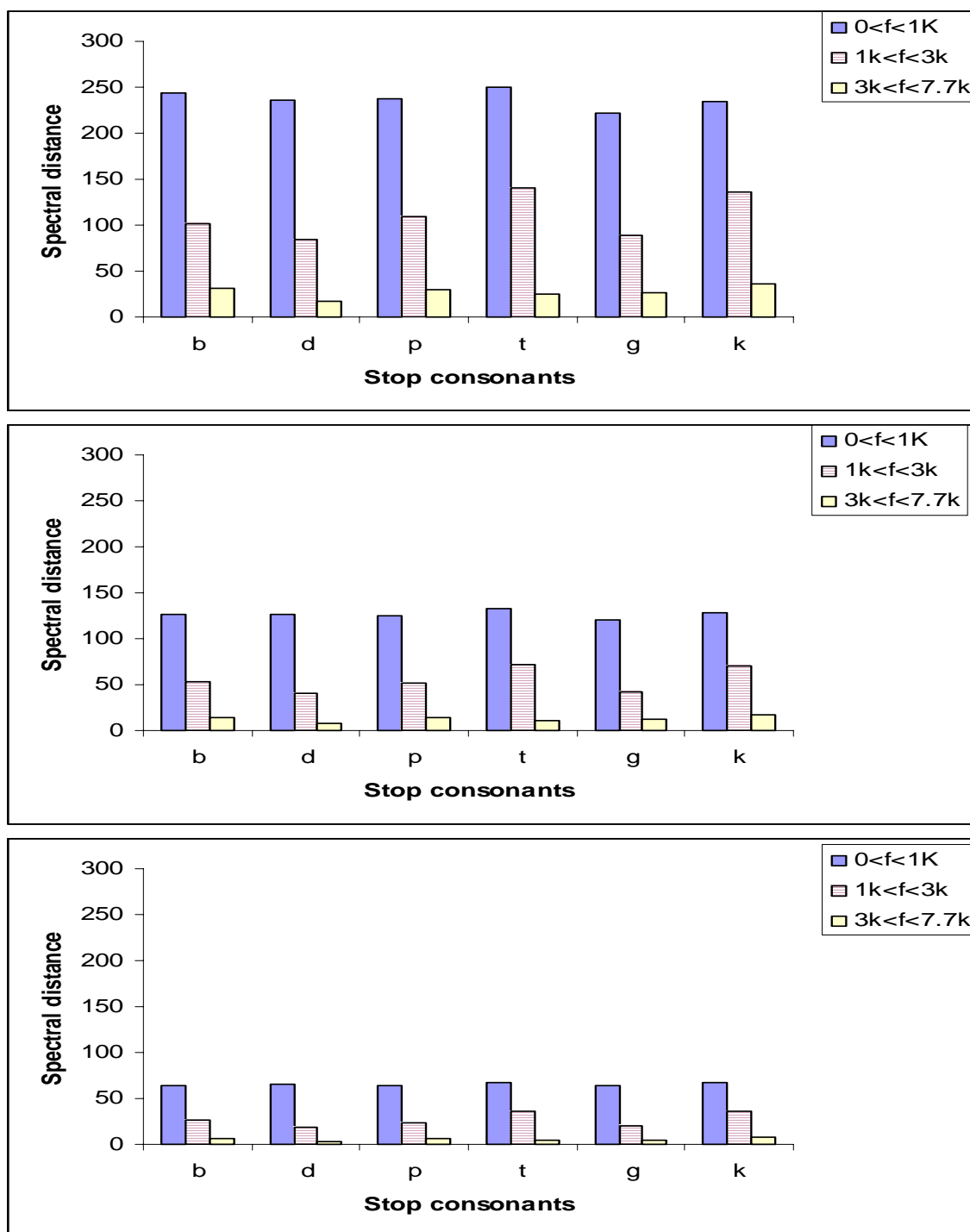
#### 4.6 Spectral distance measurement for consonants

Spectral distance measurements were also done for the stop consonants in noise at -5, 0, 5, and 10 dB SNR. Figure 4.11 shows the spectral distance plot for -5 to 10 dB speech-shaped noise added to the stop consonants. Similar plots were obtained for babble noise as shown in Figure 4.12.

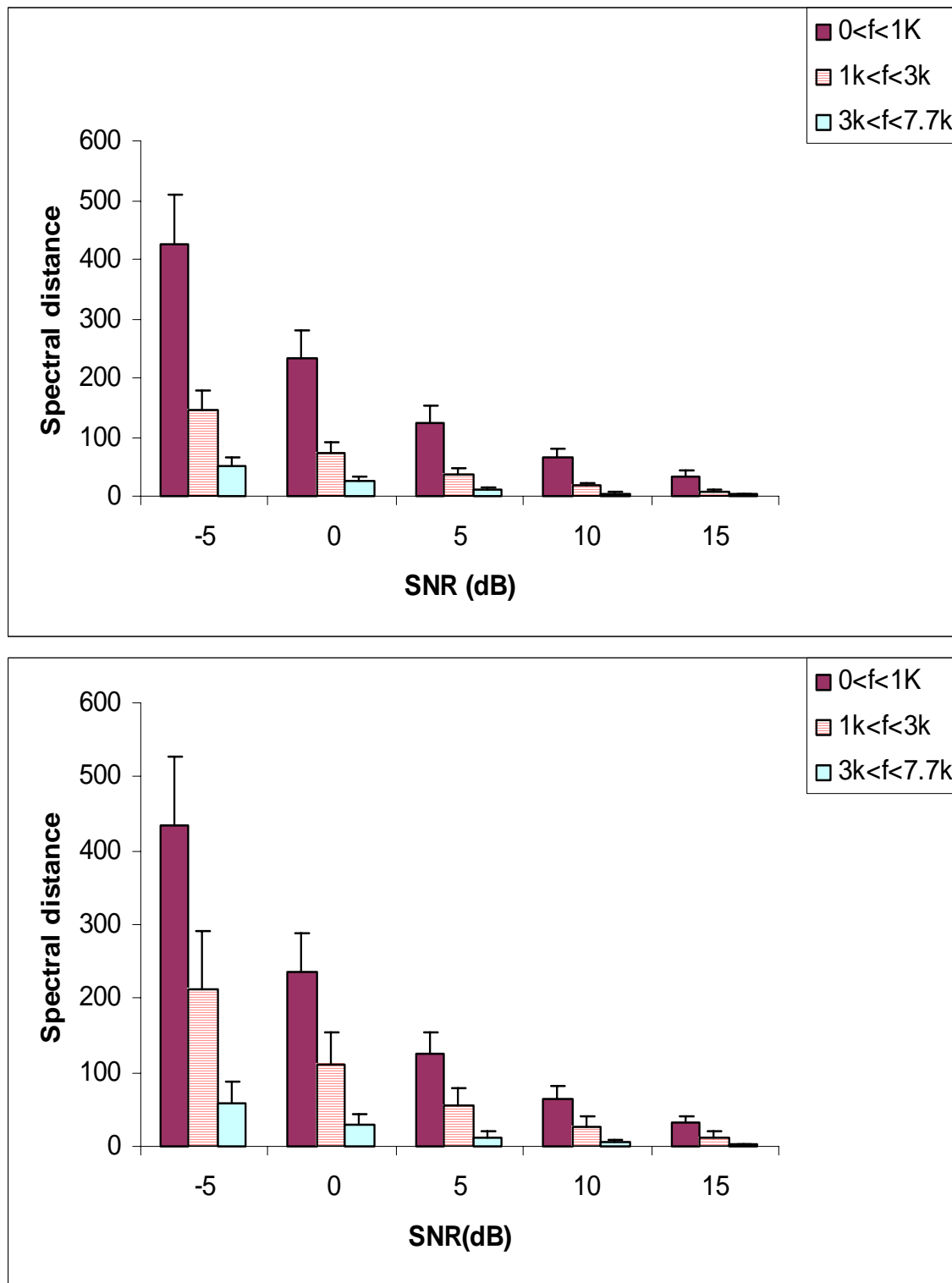
The average spectral distance measurements for the 3 frequency bands considered are shown in Figure 4.13 for speech-shaped noise (top panel) and multi-talker babble (bottom panel). Overall, the spectral distance decreased as the SNR increased. The largest spectral difference between the noisy and clean consonant spectra occurred in the low-frequency band (0-1 kHz), followed by the middle-frequency band (1-2.7 kHz) and the high-frequency band (2.7-8 kHz). This pattern was consistent for both types of noise. These



**Figure 4.11.** Spectral distance results for individual stop consonants corrupted by speech-shaped noise. (Top panel) Results for 0 dB speech-shaped noise. (Middle panel) 5 dB speech-shaped noise (Bottom panel) 10 dB speech-shaped noise.



**Figure 4.12.** Spectral distance results for individual consonants corrupted by multi-talker babble. (Top panel) Results for 0 dB multi-talker babble. (Middle panel) 5 dB multi-talker babble (Bottom panel) 10 dB multi-talker babble.



**Figure 4.13.** Average spectral distance for stop consonants. (Top panel) Spectral distance results for speech-shaped noise. (Bottom panel) Spectral distance results for multi-talker babble.

results are justified by the fact that both multi-talker babble and speech-shaped noise affect primarily the low and mid frequency regions of the spectrum.

#### **4.7 Burst frequency measurement for stop consonants**

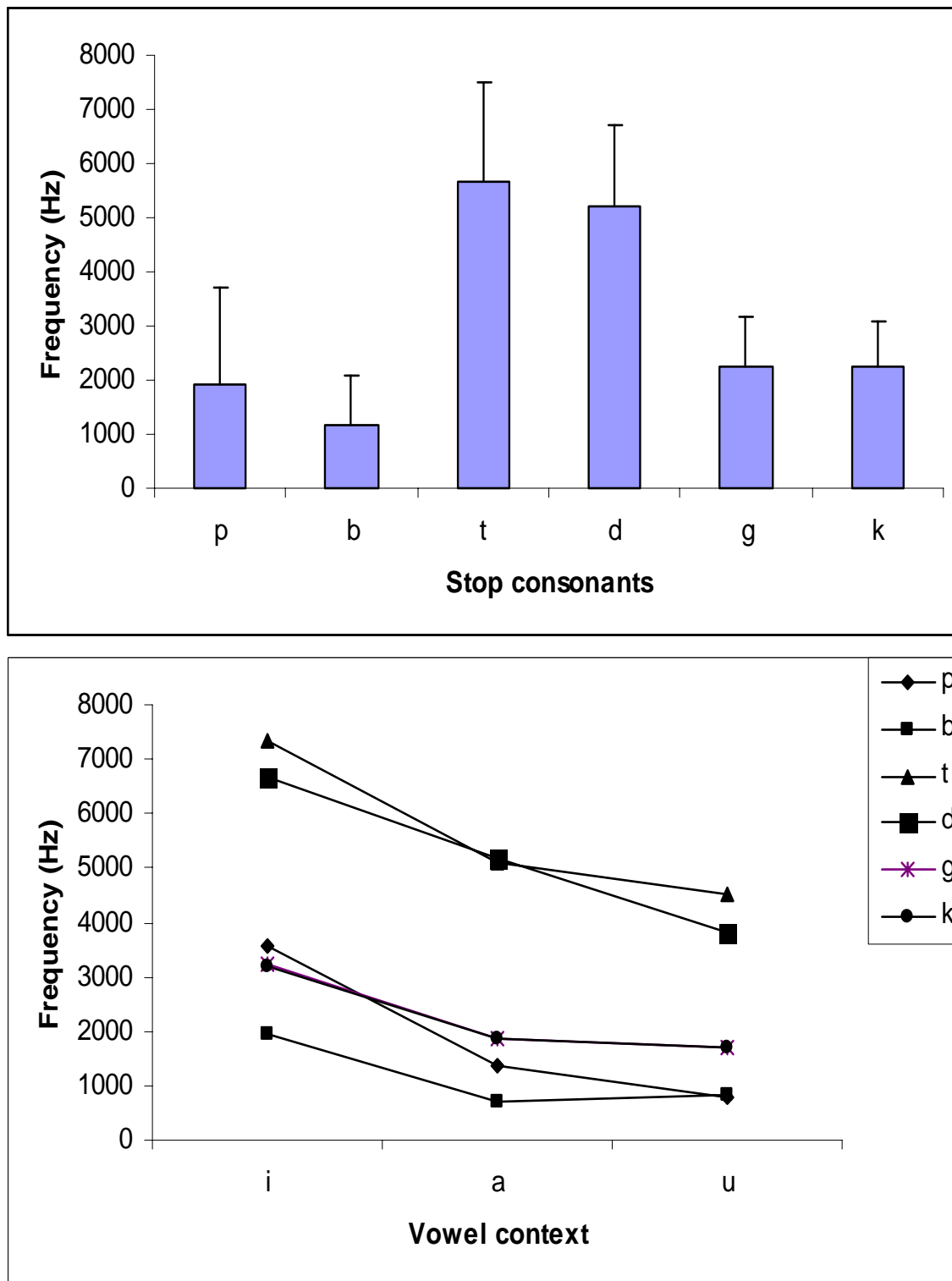
The burst frequency was computed for both the quiet and the stop consonants with noise added at -5 to 10 dB. The purpose of this analysis was to investigate the effect of the noise on the burst frequency. Figures 4.15 and Figure 4.16 show the burst frequency measurements under each SNR condition for speech-shaped and multi-talker babble.

As seen from the plots, the alveolars (/t/ and /d/) have a larger value of burst frequency than the labials (/b/ and /p/) stop consonants. This fact can be reaffirmed by examining Figure 3.9 which shows that the alveolars have higher burst frequency than those of labials. Also, velars have burst frequency higher than labials but smaller than alveolars.

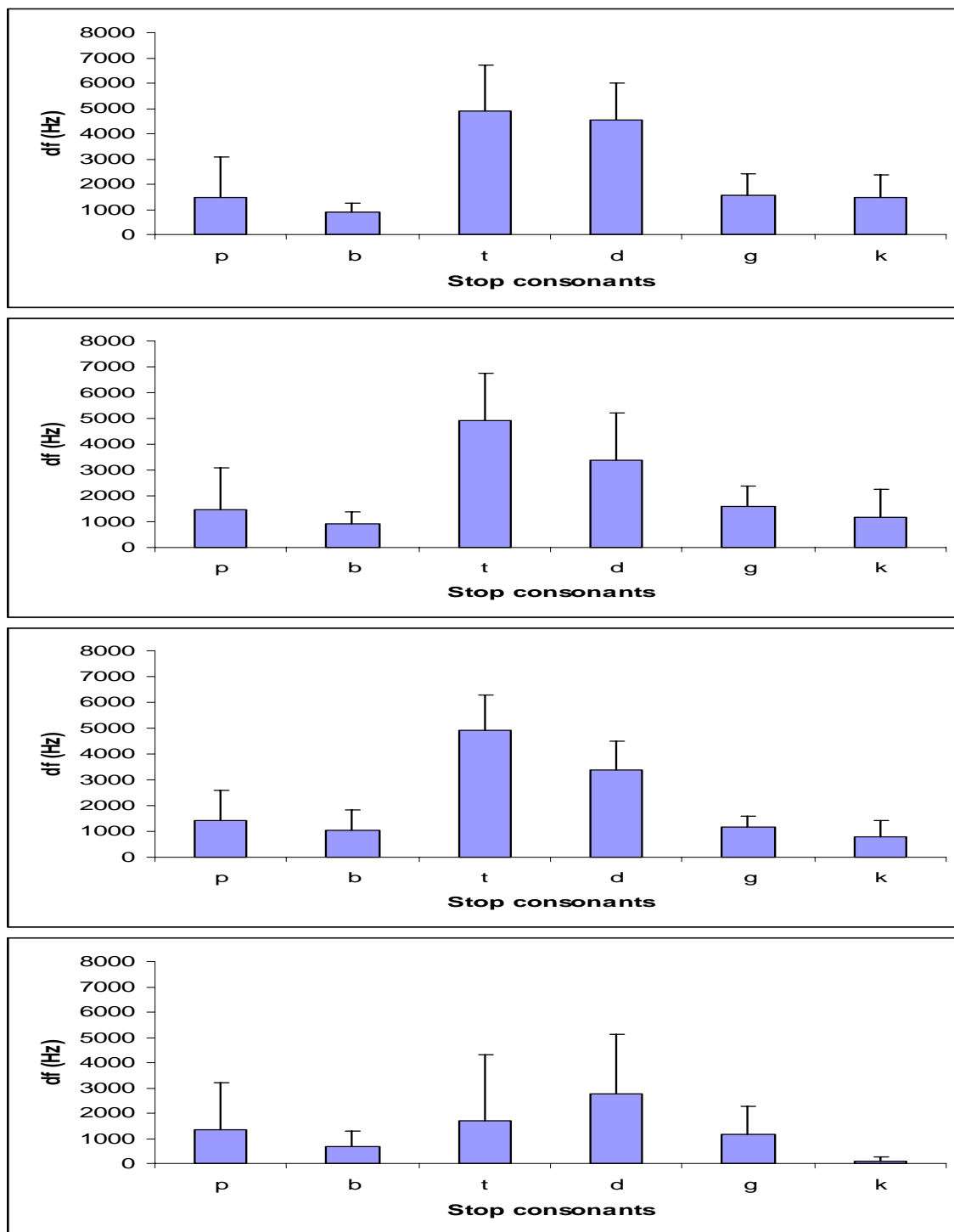
Figure 4.17 shows the deviation in burst frequency averaged across all stop consonants under different SNR and noise conditions. As seen, multi-talker babble causes a larger frequency deviation in the stop consonants than speech-shaped noise.

#### **4.8 Tilt measurement of burst frequency spectrum**

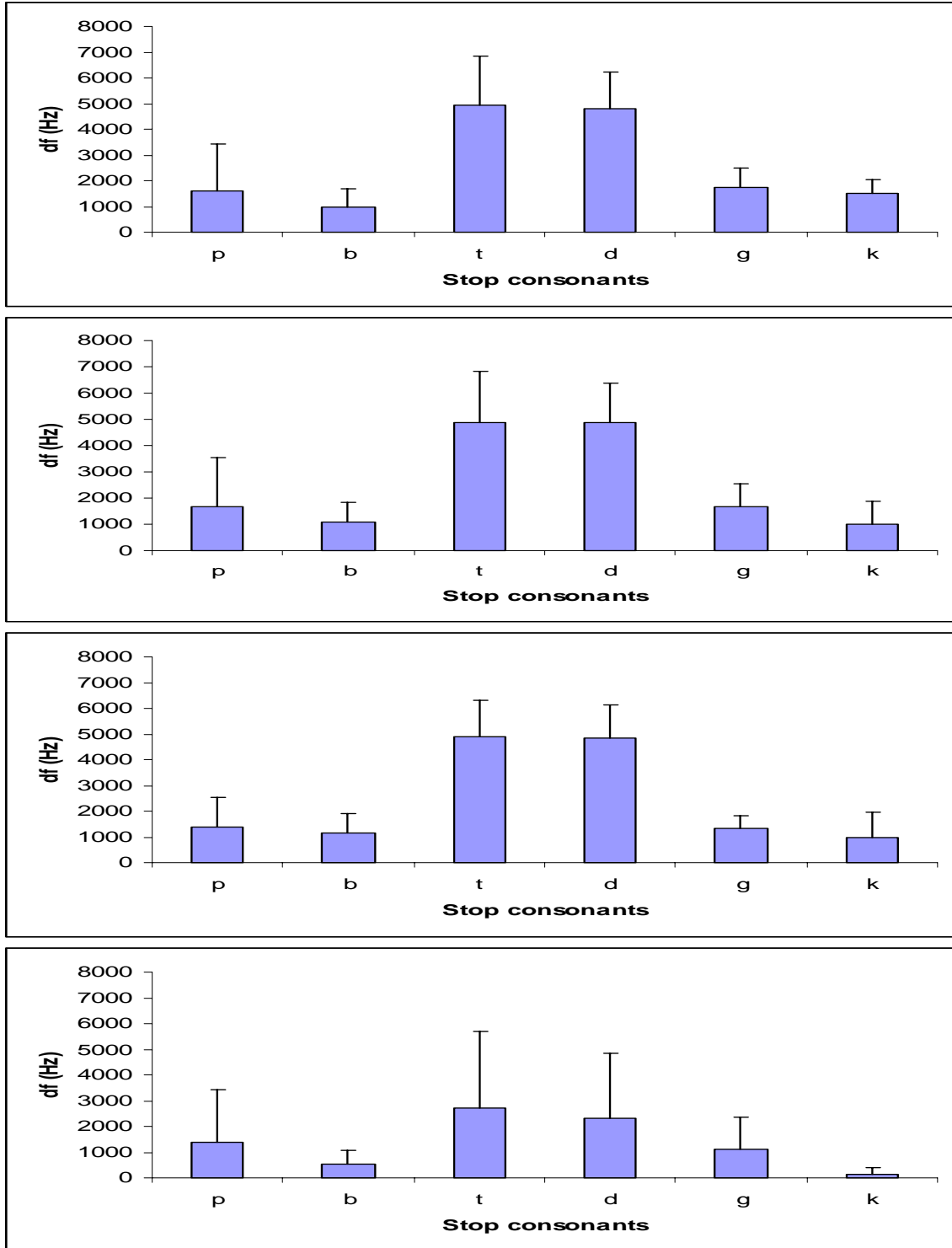
Order 4- LPC spectrum was used and the amplitude of the LPC spectrum was measured at 1000 Hz and 5000 Hz. The dB difference for the amplitude at 5000 Hz and 1000 Hz was measured and divided by the frequency interval in order to get the slope. Figure 4.18 shows the slope of the tilt at various SNRs for speech-shaped noise (top panel) and multi-talker babble (bottom panel). As the results suggest, the stop consonants /t/ and /d/ are most affected by noise at low SNR. As seen, the tilt of the stop consonants /t/ and /d/ are



**Figure 4.14.** Burst frequency for clean stop consonants. (Top panel) Average value of clean frequency for stop consonants found out considering all the vowel contexts. (Bottom panel) Averaged burst frequency for a particular vowel context.

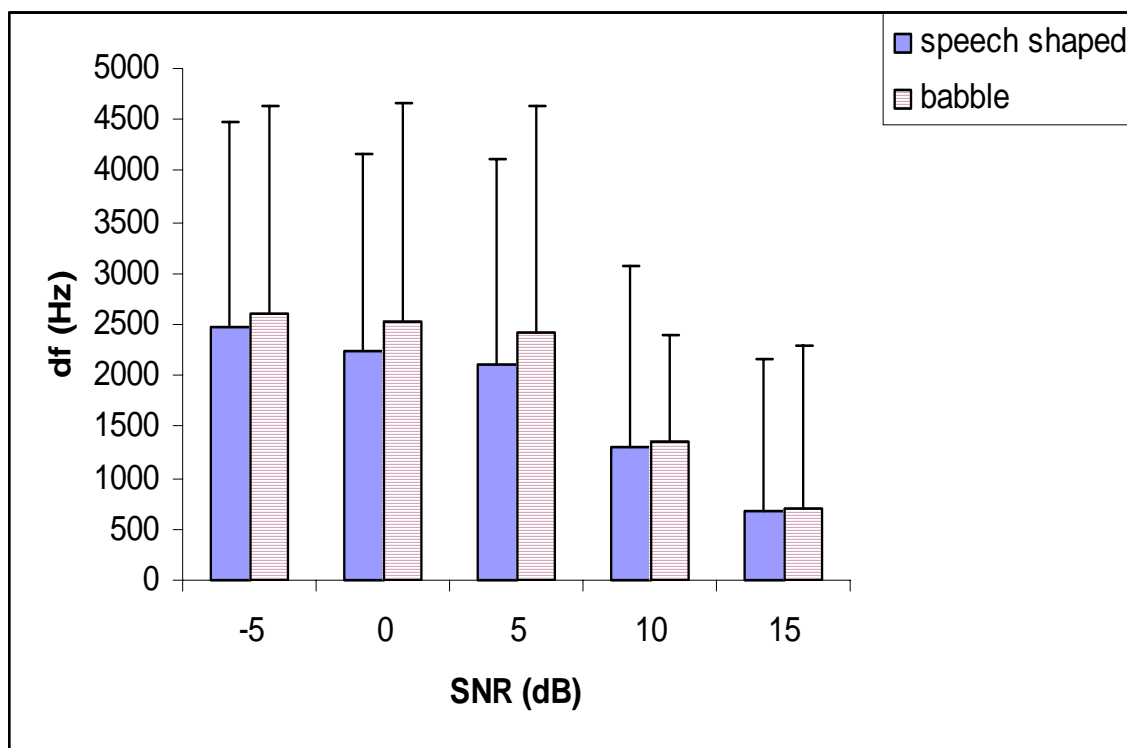


**Figure 4.15.** Burst frequency for different difference between quiet and noisy consonants for various SNR for speech-shaped noise. (Top panel) -5 dB SNR, with 0 dB SNR (Second panel from top), with 5 dB SNR (third panel from top) and 15 dB SNR (bottom panel).

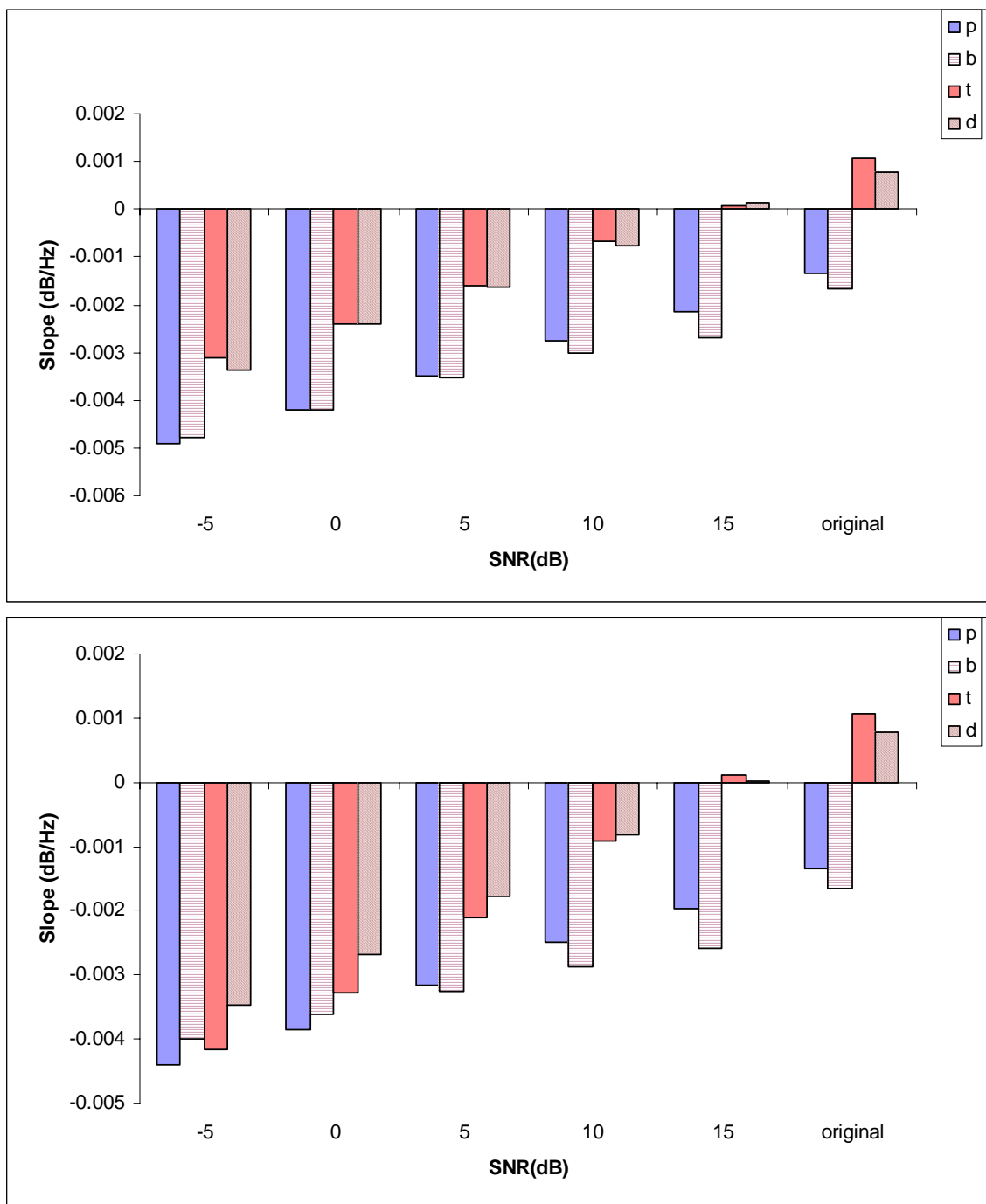


**Figure 4.16.** Burst frequency difference between quiet and noisy consonants for different SNR for multi-talker babble. (Top panel) -5 dB SNR, with 0 dB SNR (Second panel from top), with 5 dB SNR (third panel from top) and 15 dB SNR (bottom panel).

reversed. Both type of noise also affect the labial stop consonants /b/ and /p/ but the trend of the tilt for the noisy consonant remains the same.



**Figure 4.17.** Comparison of speech-shaped and multi-talker babble results for burst frequency difference measurements.



**Figure 4.18.** Measurement of tilt in the spectrum for different SNR. (Top panel) Slope measurements for different SNR for speech-shaped noise. (Bottom panel) Slope measurement for different SNR for multi-talker babble.

## CHAPTER FIVE

### SUMMARY AND CONCLUSIONS

The work in this thesis focused on quantifying the effects of noise on the frequency spectrum of vowels and consonants. Acoustic measures like spectral contrast, spectral distance, formant frequency estimation and deviation in formant frequency were employed for quantifying the effect of noise on vowels whereas measures such as spectral distance, spectral tilt, and burst frequency were employed for stop consonants. The results established that the type of colored noise employed in this thesis had a non-uniform effect on the frequency spectrum of the vowels and consonants.

The major conclusions of this thesis are:

**(a) Spectral contrast reduction**

It was shown that noise tends to flatten the spectrum and reduce the spectral contrast by as much as 20 dB for low-SNR conditions. This suggests that speech enhancement algorithms should incorporate a post-processing stage to spectrally enhance the speech thereby compensating for the reduced spectral contrast. Such techniques are in fact used in speech coders (Chen and Gersho, 1987).

**(b) Effect of number of channels on spectral contrast**

It was shown that for a constant SNR, as the number of channels increased, the vowel spectral contrast increased. This indicated that increasing the number of channels could somehow compensate for the low SNR.

**(c) Non-uniform spectral effect on vowels and consonants**

Results of spectral distance establish that the various bands of the spectrum are affected differently by noise, some bands to a smaller degree and some to a larger degree. The low-frequency band (0-1 kHz) is affected the most, followed by the mid-frequency band (0-2.7 kHz), followed by the high-frequency band (3-8 kHz). This fact needs to be exploited by speech enhancement algorithms in order to improve the quality of speech. Also, this finding suggests a multi-band approach to speech enhancement where each band is treated differently and probably weighted differently. Such an approach was proposed recently by Kamath and Loizou (2002), in the context of spectral subtraction. Although the magnitude of the effect differed between multi-talker babble and speech-shaped noise, both types of noise showed a similar trend.

**(d) Formant frequency estimation**

Since the formants F1 and F2 played a significant role in vowel perception, it was desirable to find out the effect of noise as well as SNR on the detection of the first two formant frequencies.

For the low SNR conditions, the first formant (F1) was reliably detected more often compared to the second formant (F2). In speech-shaped noise at  $-5$  dB S/N, F1 was detected 64% of time, while F2 was detected 48% of the time. The fact that F2 was detected less often than F1 is consistent with the spectral distance measurements which showed that the multi-talker babble affects the F2 region of the spectrum to a larger degree than the speech-shaped noise does. Both F1 and F2 formants were reliably detected more than 50% of the time only for SNR values of 5 dB and higher.

**(e) Burst frequency measurement for stops**

A study of how the type of noise and SNR affects the burst frequency was done by measuring the frequency deviation between clean and noisy consonants. The results from the study showed that the largest frequency deviation occurred for alveolars, followed by velars and labials. This result was true for both types of noise. Also, as the SNR increased the deviation in the burst frequency decreased. Similar trend was observed for both types of noises.

**(f) Spectral tilt measurement for stops**

The measurements for the tilt measurement showed that the alveolars /t/ and /d/ are affected to a larger extent compared to the labials /b/ and /p/. The results showed that while the labials retained the spectral tilt even at very low SNR (although the magnitude of the slope differed), the tilt of the spectrum for alveolars /t/ and /d/ were reversed even at a relatively high SNR of 10 dB. Hence, we can conclude that the noise employed had a detrimental effect on alveolars.

Further research work can be conducted in incorporating the results from this thesis in noise reduction algorithms. For instance, a sub-band approach can be employed in most noise reduction algorithms. For instance, more aggressive subtraction can be done in the lower frequency range than in the upper frequency range. Similarly, future studies investigating the effect of employing a post processing stage in order to compensate for the reduction in the spectral contrast can be done.

## REFERENCES

- [1] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procs.*, pp. 208-211, Apr. 1979.
- [2] S. Blumstein and K. Stevens, "Acoustic invariance in speech production," *J. Acoust. Soc. Am.*, vol. 66, pp. 1001-1017, 1979.
- [3] S. Blumstein and K. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Am.*, vol. 67, pp. 648-662, 1980.
- [4] S. Blumstein, E. Issac and J. Mertus, "The role of gross spectral shape as a perceptual cue to place of articulation", *J. Acoust. Soc. Am.*, vol. 72, pp. 43-50, 1982.
- [5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113-120, Apr. 1979.
- [6] J. Chen and A. Gersho, "Real-time VAPC speech coding at 4800 bits/sec with adaptive postfiltering," *Proc. ICASSP*, pp. 2185-2188, 1987.
- [7] Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *ICASSP*, vol. 2, pp. 961-964, Apr. 1991.
- [8] F. Cooper, P. Delattre, A. Liberman, J. Borst and L. Gerstman, "Some experiment on perception of synthetic speech sounds", *J. Acoust. Soc. Am.*, vol. 24, pp. 597-606, 1952.

- [9] P. Delattre, A. Liberman, F. Cooper and L. Gerstman, "An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns," *Word*, vol 8, pp. 195-210, 1952.
- [10] J. Deller Jr, J. Proakis and J. Hansen, *Discrete-time processing of speech signals*, Macmillan, 1993.
- [11] M. Dorman, M. Studdert-Kennedy and L. Raphael, "Stop consonant recognition: Release bursts and formant transitions as functionally equivalent context-dependent cues," *Percept. Psychophys.*, vol. 22, pp. 109-122, 1977.
- [12] M. Dorman and P. Loizou, "Relative spectral change and formant transitions as cues to labial an alveolar place of articulation," *J. Acoust. Soc. Am.*, vol. 100, pp. 3825-3830, 1996.
- [13] G.Fant, *Acoustic Theory of Speech Production*, 's-Gravenhage, The Netherlands: Mounton and Co., 1960.
- [14] J. Flanagan, "A difference limens for vowel formant frequency," *J. Acoust. Soc. Am.*, vol. 27, pp. 288-291, 1955.
- [15] B. Gold and N. Morgan, *Speech and audio signal processing*, Wiley, 2000.
- [16] J. Hawks, "Difference limens for formant patterns of vowel sounds," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1074-1084, 1994.
- [17] J. Hillenbrand and R. Gayvert, "Identification of steady-state vowels synthesized from the Peterson and Barney measurements," *J. Acoust. Soc. Am.*, vol. 94, pp. 668-674, 1993.

- [18] 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.
- [19] T. Hirahara and H. Kato, "The effect of F0 on vowel identification," *Speech Perception, Production, and Linguistic Structure*, eds Y.Tokhura, E. Vatikiotis-Bateson, and Y. Sagisaka Tokyo: Ohmusha Ltd., 1992.
- [20] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. ICASSP*, Orlando, FL, 2002.
- [21] D. Kewley-Port, "Time varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 73, pp. 322-335, 1983.
- [22] D. Kewley-Port, "Thresholds for formant-frequency discrimination of vowels in consonantal context," *J. Acoust. Soc. Am.*, vol. 97, pp. 3139-3146, 1995.
- [23] D. Kewley-Port, "Psychophysical studies of vowel formants," *Proceedings of the Keele Workshop on the Auditory Basis of Speech Production*, eds. W.A. Ainsworth and S. Greenberg, 148-153, 1996.
- [24] D. Kewley-Port, "Detection thresholds for isolated vowels," *J. Acoust. Soc. Am.*, vol. 89, pp. 820-829, 1991.
- [25] D. Kewley-Port, X. Li, Y. Zheng, and A. Neel, "Fundamental frequency effects on the thresholds for vowel formant discrimination," *J. Acoust. Soc. Am.*, vol. 100, pp. 2462-2470, 1996.
- [26] D. Kewley-Port and C. Watson, "Formant frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.*, vol. 95, pp. 485-496, 1994.

- [27] D. Kewley-Port and Y. Zheng, "Vowel formant discrimination in ordinary listening conditions I," *J. Acoust. Soc. Am.*, vol. 100, pp. 2689, 1996.
- [28] A. Lahiri, L. Gerwirth and S. Blumstein, "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants. Evidence from a cross-language study," *J. Acoust. Soc. Am.*, vol. 76, pp. 391-404, 1984.
- [29] M. Leek, M. Dorman and Q. Summerfield, "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 81, pp. 148-154, 1987.
- [30] P. Loizou and O. Poroy, "Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners", *J. Acoust. Soc. Am.*, vol. 110, pp. 1619-1627, 2001.
- [31] J. Miller, "Auditory-perceptual interpretation of the vowel", *J. Acoust. Soc. Am.*, vol. 85, pp. 2114-2134, 1989.
- [32] T. Nearey and P. Assman, "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.*, vol. 80, pp. 1279-1308, 1986.
- [33] T. Nearey, "Static, dynamic and relational properties in vowel perception," *J. Acoust. Soc. Am.*, vol. 85, pp. 2088-2113, 1989.
- [34] A. Nebalek and P. Dagenais, "Vowel errors in noise and in reverberation by hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 80, no.3, pp. 741-748, 1986.
- [35] A. Nebalek, "Identification of vowels in quiet, noise and reverberation: Relationships with age and hearing loss," *J. Acoust. Soc. Am.*, vol. 84, no. 2, pp. 476-484, 1988.

- [36] A. Neel and D. Kewley-Port, "Formant movements and duration cues in the identification of vowels," *J. Acoust. Soc. Am.*, vol. 96, pp. 3284, 1994.
- [37] A. Neel and D. Kewley-Port, "Dynamic cues in vowel identification: A training study," *J. Acoust. Soc. Am.*, vol. 99, pp. 2593, 1996.
- [38] A. Neel, "Factors affecting vowel identification in hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 102, No. 4, Pt. 2, 3095, presented at the 134<sup>th</sup> meeting of the Acoustical Society of America, San Diego, California, November, 1997.
- [39] O' Shaughnessy, "*Speech communications. Human and machine*," IEEE Press, 2000.
- [40] G. Peterson and H. Barney. "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, pp. 175-184, 1952.
- [41] J. Pickett, "Perception of vowels heard in noises of various spectra," *J. Acoust. Soc. Am.*, vol. 29, no.3, pp. 613-620, 1957.
- [42] D. Pisoni, "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," *J. Acoust. Soc. Am.*, vol. 61, pp. 1352-1361, 1977.
- [43] S. Quackenbush, T. Barnwell III and M. Clements, "*Objective measures of speech quality*," Prentice Hall, 1988.
- [44] R. Shannon, A. Jensvold, M. Padilla, M. Robert, and X. Wang, "Consonant recording for speech testing", *J. Acoust. Soc. Am.*, vol. 106 pp. L71-L74, 1999.
- [45] K. Stevens and S. Blumstein, "Invariant cues for the place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 64, pp. 1358-1368, 1978.

- [46] W. Strange, "Evolving theories of vowel perception," *J. Acoust. Soc. Am.*, vol. 85, no.5, pp. 2081-2081, 1989.
- [47] W. Strange, J.J. Jenkins, and T.L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, vol. 74, pp. 695-705, 1983.
- [48] A. Syrdal and H. Gopal. "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.*, vol. 79, no.4, pp. 1086-1100, 1986.
- [49] D. Van Tasell, D. Fabry and L. Thibodeau, "Vowel identification and vowel masking patterns of hearing-impaired subjects," *J. Acoust. Soc. Am.*, vol. 81, pp. 1586-1597, 1987.
- [50] E. Zwicker and H. Fastl., *Psychoacoustics: Facts and Models*, Springer Verlag: Berlin, 1999.
- [51] E. Zwicker and H. Fastl., *Psychoacoustics: Facts and Models*, Springer Verlag: Berlin, 1999.

## **VITA**

Gaurang Parikh completed his Bachelors in Engineering from University of Mumbai with the highest honors where he stood first among 566 students in the department of Electronics and Telecommunications in the year 1999. He joined CMC (India) Ltd., where he was responsible for supervising and maintaining VAX/ VMS mainframe system for on-line ticket reservations for India Railways.

He was admitted to the Masters program at the University of Texas at Dallas, Richardson in the Department of Telecommunication Engineering in August 2000. He worked as a Research Assistant under the supervision of Dr. Loizou, in the Electrical Engineering Department at the University of Texas at Dallas till August 2002. His research interests are in the field of speech processing for cochlear implants.