DICHOTIC SPEECH RECOGNITION: ACOUSTIC AND ELECTRIC HEARING

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Philipos C Loizou, Chair

Dr. Issa M S Panahi

Dr. Mohammad Saquib

Dedicated to

my grand parents and parents

DICHOTIC SPEECH RECOGNITION: ACOUSTIC AND ELECTRIC HEARING

by

ARUNVIJAY MANI, B.E. in EIE

DISSERTATION

Presented to the Faculty of

The University Of Texas At Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

MAJOR IN TELECOMMUNICATIONS

THE UNIVERSITY OF TEXAS AT DALLAS

May 2004

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper, or papers, submitted for publication. The dissertation must still conform to all other requirements explained in the “Guide for Preparation of Master’s Theses, Doctoral Dissertations, and Doctor of Chemistry Practica Reports at The University of Texas at Dallas. It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student’s contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

ACKNOWLEDGEMENTS

I would like to thank my adviser, Dr. Philipos Loizou, for his guidance in my studies and my work. He has offered me many helpful suggestions on conducting research and writing technical documents.

I would also like to thank Dr. Mohammad Saquib and Dr. Issa Panahi for their valuable feedback on this manuscript.

I also thank Dr. Michael Dorman of Arizona State University for his useful comments on my work. I would like to thanks Tony Spahr for his help with the TIMIT database files.I would like to grab this opportunity to express my deepest gratitude to Dr. Arthur Lobo of University of Texas at Dallas for his constant encouragement and useful comments on my work.

I like to thank Dr.Vandali of Cooperative Research Center for Cochlear Implant and Hearing Aid Innovation (CRC), for permitting to use some contents from their documentation of SPEAR3 cochlear implant.

I also thank Prasanna Krishnamoorthy, Kalairajah Narayanan for helping me format the thesis. And finally I take this chance to thank all of my friends for being my driving force and standing by me through thick and thin.

DICHOTIC SPEECH RECOGNITION: ACOUSTIC AND ELECTRIC HEARING

Publication No. ____________________

ArunVijay Mani, M.S.
The University Of Texas At Dallas, 2004

Supervising Professor: Dr. Philipos C Loizou

It is generally accepted that the fusion of two speech signals presented dichotically is affected by the relative onset time. This study investigated the hypothesis that spectral resolution might be an additional factor influencing spectral fusion when the spectral information is split and presented dichotically to the two ears. Two different methods of splitting the spectral information were investigated. In the first method, the odd-index channels were presented to one ear and the even-index channels to the other ear. In the second method the lower frequency channels were presented to one ear and the high frequency channels to the other ear. The experiments were conducted with both normal hearing listeners and bilateral cochlear implant listeners. Results with normal hearing listeners indicated that spectral resolution did affect spectral fusion. Results with bilateral cochlear implant users indicated that subjects were able to fuse information presented to the two ears accurately in quiet but not in noise.

TABLE OF CONTENTS

PREFACE v

ACKNOWLEDGEMENTS vi

ABSTRACT vii

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 LITERATURE REVIEW 3

2.1 Introduction 3

2.2 Fusion of Speech 3

2.3 Effect of Intensity Differences on Fusion 6

2.4 Effect of Fundamental Frequency 9

2.5 Dichotic Interleaved Presentation of Speech Signals 10

2.6 Summary 12

CHAPTER 3 DICHOTIC PRESENTATION FOR NORMAL LISTENERS 13

3.1 Chapter Outline 13

3.2 Subjects 13

3.3 Speech Materials 14

3.4 Signal Processing 14

3.5 Procedure 16

3.6 Results 18

3.6.1 Sentences 18

3.6.2 Vowels 20

3.6.3 Consonants 22
3.7 Discussions 24

CHAPTER 4 DICHOTIC SPEECH RECOGNITION BY BILATERAL COCHLEAR IMPLANTS LISTENERS 32
4.1 Chapter Outline 32
4.2 SPEAR3 Research Processor 32
4.2.1 Introduction 32
4.2.2 Functional Description 33
4.2.3 SPEAR3 Speech Processing System 35
4.3 Subjects 36
4.4 Speech Materials 37
4.5 Experimental Setup 37
4.6 Procedure 38
4.7 Results 39
4.7.1 Results in Quiet 39
4.7.2 Results in Noise 41
4.8 Effects of Beamforming on Intelligibility 43
4.8.1 Introduction 43
4.8.2 Head Related Transfer Function 43
4.9 Beamforming Experiments 45
4.9.1 Generating the Noisy Binaural Speech 45
4.9.2 Beamforming 46
4.10 Results from Beamforming 49
4.11 Discussion 49

CHAPTER 5 SUMMARY AND CONCLUSIONS 51

REFERENCES 54

VITA

LIST OF TABLES

3.1 The 3-dB cutoff frequencies (Hz) of the bandpass filters used in this study. FL and FH indicate the low and high cutoff frequencies respectively of the bandpass filters. 15

4.1 Biographical data of subjects. 37

LIST OF FIGURES

3.1 Mean identification of sentences (scored in percent words correct), vowels and consonants as a fraction of number of channels and presentation mode (monaural vs. dichotic). Error bars indicate standard errors of the mean. 19

3.2 (Top panel) Mean speech recognition obtained when the odd- and even-index channels were presented dichotically (hatched bars), the odd-channels were presented to the left ear only (diagonally filled bars) and the even-channels were presented to the right ear only (dark bars). (Bottom panel). Mean vowel and consonant recognition obtained when the low- and high-frequency channels were presented dichotically (hatched bars), the low-frequency channels were presented to the left ear only (diagonally filled bars) and the high-frequency channels were presented to the right ear only (dark bars). Speech materials were processed through 12 channels. Error bars indicate standard errors of the mean. 21

3.3 Mean identification of the consonant features manner, place and voicing as a function of number of channels and presentation mode (monaural vs. dichotic). Error bars indicate standard errors of the mean. 23

3.4 Individual subject performance for speech materials processed through 6 channels. 25

4.1 SPEAR3 Functional Diagram 33

4.2 Seed Speak Client Map Settings Window 36

4.3 Mean scores for the diotic and the two dichotic condition in quiet. . . 39

4.4 Mean speech recognition obtained when the odd- and even-index channels are presented dichotically, the odd-channels are presented to the left ear only and the even-channels are presented to the right ear only. All the tests are conducted in quiet. 40

4.5 Mean speech recognition obtained when the low- and high-index channels are presented dichotically, the low frequency-channels are presented to the left ear only and the high frequency-channels are presented to the right ear only. All the tests are conducted in quiet. 40

4.6 Mean score for the diotic and two dichotic conditions in noise. 41

4.7 Mean speech recognition obtained when the odd- and even-index channels are presented dichotically, the odd-channels are presented to the left ear only and the even-channels are presented to the right ear only. All the tests are conducted in noise with SNR of +10dB. 42

4.8 Mean speech recognition obtained when the low- and high-index channels are presented dichotically, the low frequency-channels are presented to the left ear only and the high frequency-channels are presented to the right ear only. All the tests are conducted in noise with SNR of +10dB. 42

4.9 Head Shadow Effect. 44

4.10 Beamforming using Jim Griffiths algorithm. 46

4.11 Individual scores obtained when unprocessed speech presented binaurally, beamformed output, beamformed output in one ear and the unprocessed speech in the other ear. 48

CHAPTER 1

INTRODUCTION

Binaural hearing enable us to localize the source of sounds and understand speech in noisy conditions. Early research on bilateral implants predicted that bilateral implantation would more closely approximate the binaural hearing, because it provides additional acoustic information. Early results also showed that bilateral implants improve speech recognition in noise when compared to the unilateral implants [1].

The question of whether implanting both ears is better than implanting one ear continues to be a focus of researches in bionic ear development. Efforts were made in the past to examine whether bilateral implants exhibit the binaural advantage as normal listeners. Of major importance is the understanding of how information is integrated centrally when presented to the two ears. Research has shown that the fusion of information presented to the two ears is a determining factor in obtaining the binaural advantage.

Fusion of spectral information becomes critical when the information is presented dichotically to the bilateral implants. In dichotic presentation mode, the information given to the two ears are different.

Of the three factors investigated in the literature, the relative onset-timing difference seemed to have the largest effect on dichotic speech perception. The recent introduction of bilateral cochlear implants spurred the question of whether cochlear implant listeners would able to fuse speech information presented dichotically. The aim of this study is to investigate the hypothesis that the spectral resolution might be an additional factor influencing fusion when the speech signal is presented dichotically. This study also investigates the potential benefits of dichotic electrical simulation, namely reduction of channel interaction, since one can simulate the electrodes alternately across the two ears, reduction in power consumption, etc.

To examine the effect of spectral resolution on dichotic speech recognition, noisy speech was processed into small numbers (6-12) of channels and presented to normal-hearing listeners dichotically. Two different conditions were considered. In the first condition, low frequency information was presented to one ear and high frequency information was presented to the other ear. In the second condition, the frequency information was interleaved between the two ears with the odd-index frequency channels fed to one ear and the even-index channels fed to the other ear. The issue is whether spectral fusion is affected by: (1) poor spectral resolution, or/and (2) the way spectral information is split (low/high vs. interleaved) and presented to the two ears. We hypothesize that both spectral resolution and the type of spectral information presented to the two ears will affect spectral fusion.

The thesis is organized as follows. Chapter 2 presents a review of related literature in the field of dichotic speech perception and fusion. In Chapter 3, the experimental procedures and the results with the normal hearing listener are presented. Chapter 4 presents the experiments and the results with binaural cochlear implants. Chapter 4 also discusses the effects of beamforming on the speech recognition. Chapter 5 summarizes the work performed and presents the conclusions.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of related literatures starting from the basic introduction to dichotic listening to description of the factors affecting dichotic listening.

Section 2.2 reviews the literature dealing with the fusion of speech, section 2.3 presents the literature dealing with the effect of intensity. Section 2.4 reviews the effect of fundamental frequency. Section 2.5 reviews the interleaved presentation mode.

2.2 Fusion of Speech

Cherry [2] analyzed the recognition of speech by varying the speech signal reaching the two ears. In one condition the listeners were presented with a mixture of speech signals to both ears and in the second condition, one speech signal in the mixture was given to one ear while other speech signal was presented to the other ear.

In the first condition the two separate messages were recorded together and presented monophonically, and the listeners were asked to report one of the messages. The subjects were allowed to repeat the messages till they were sure about the message they want to identify. The messages that were mixed were selected such that they were highly correlated to each other. The subjects reported great difficulty in recognizing the particular sentences. In order to aid the subjects, paper and pencil was provided so that they can take notes while listening. As a variation to this condition, the messages spoken by the same speaker was recorded together and presented to the subjects. Cherry [2] reported that the subjects could get hidden message when they get two or three words contained in the messages. They seemed to fill up the rest of the message from their past training.

In the second condition the messages were presented in such a way that one message was presented in one ear while the other message was presented in the other ear. The subjects had no problem in identifying the required message and reject the other message completely. The subjects had no problem in repeating the identified message. But at the same time the subjects did not have a clue about the rejected message.

The condition was varied by recording the same message in both the ears and the onset time on the ears is varied. Cherry [2] reported that the recognition of the message presented depends on the magnitude of the difference in the onset time of the two ears. The condition was further complicated by switching the message between the ears. The subjects were altered to shift their concentration from one ear to the other provided the shift is slow. But their concentration deteriorated with the increase in the speed of the switching. This was due to the fact that the noise level masking the message was extremely low in case of the slow shifting, while the noise level increased with the increase in speed of shifting.

Introduction of dichotic presentation by Cherry [2] and its application in cocktail party effect motivated researchers to investigate dichotic listening and the factors affecting the efficiency of dichotic listening.

Broadbent *et al.* [3] examined the problem raised by the simple place theory of hearing. For identifying a vowel in the presence of other sounds, required that all the formants of the vowel be detected as such and not classified with other sounds. Broadbent *et al.* [3] demonstrated this concept by combining the word bid and bird. The vowels presented in the words had their formants very close to each other, which makes the task of grouping the formants difficult. So the factor which was critical for the grouping of sounds was the envelope of the waveform of these sounds.

The basic variation in the recording was that in one condition all the formants were recorded together and in both tracks, while in the other condition the first formant was in one track and the second and third formant on the other track. In another variation the larynx generator frequency used to generate the formants were varied. The formant presentation to the subjects followed a specific order, the first formant was given to the left ear while the second formant was given to the right ear. This was followed for half the part of the experiment and for the rest of the experiment the pattern was reversed.

Broadbent *et al.* [3] reported that the formants fuse even when one formant was given to one ear and the other formant to the opposite ear. But when the larynx generator frequency varied, the subjects heard two different sounds. Broadbent *et al.* [3] also reported that envelope variation had no effect on the fusion when the formants were given to the same ear.

Broadbent *et al.* [3] concluded that the sustained formants of different frequencies were heard as coming from the same place even though one was presented in one ear and the other one was presented to the other ear as long as the envelope frequency was the same in both ears. When this frequency differed between the ears, the sounds seemed to be coming from two different places.

2.3 Effect of Intensity Differences on Fusion

Rand [4] investigated the dichotic method of listening using synthetic CV syllables. The experiment involved presenting formant F1 in one ear and the formants F2 and F3 in the other ear. Dichotic listening was analyzed by varying the intensity of F2 and F3. The syllables [ba,da,ga] were used as test materials. These syllables were synthesized using the Haskins Laboratories parallel resonance synthesizer. The syllables were recorded in a tape for both binaural and dichotic conditions.

In the dichotic condition, the intensity of F1 and F2 were recorded in six levels of attenuation ranging from 0 to 50 dB. For binaural listening the six levels of attenuation ranged from 0 to 30 dB. Calibration signal which was a sustained [a], a vowel was recorded on both channels for both conditions.

The experiments were conducted with four subjects who were asked to identify the consonant played. The result proved that both dichotic and binaural condition yields 100% performance for small attenuation and decreases at higher attenuations. Rand [4] reported that the dichotic presentations performance remained above the chance level for higher attenuation when compared to the binaural conditions. Approximately 20dB of attenuation separated the two curves indicating a massive release from masking, thus proving that the possible masking which occurred in the normal binaural mode could be eliminated when signal were presented dichotically.

As an addendum to the above experiment, the dichotic and binaural comparison was performed by recording the formant transition alone in one ear and the rest of the formants in the other ear. The result suggested that the binaural presentation of the main experiment produced roughly 5 dB greater masking than binaural presentation of the modified experiment. In contrast the dichotic presentation was uniform in both cases. Rand [4] concluded that the dichotic presentation mightbe advantageous to increase the intelligibility under poor SNR.

Cutting [5] examined six different fusions and investigated its robustness against variation in three parameters, namely relative onset time of the two opposite ear stimuli, their relative intensity and their relative fundamental frequency. Among these six types of fusion sound localization was the most common type and it was used as the reference for the other fusions. Cutting [5] analyzed the spectral fusion which was first reported by Broadbent *et al.* [3].

The experiments were conducted could be broadly divided into four parts. In the first part the onset time was varied and the robustness of the fusion was examined. In the second part the relative intensity was varied and the robustness of the fusions was evaluated. In the third part the fundamental frequency was varied while in the forth part the listeners were asked to report number of items heard in each condition. For the spectral fusion, syllables /ba/, /da/, /ga/ were used. The first formant of these syllables was recorded in one channel while the second formant was recorded in the second channel. For the variation in the onset timing, seven lead times were selected: 0, 10, 20, 30, 40, 80 and 160 msecs. A sequence of 72 items was recorded and the listeners were asked to write down the initial consonant that they heard clearly. Cutting [5] reported that in case of the spectral fusion, the fusion probability decreased with the increase in onset time. As the onset time increased the listeners reported all the items as /ba/ as the first formant of all the items sounds like /ba/ when presented alone.

In the second part the relative intensities were varied. One of the stimuli was maintained at 80 dB SPL while in the other channel the intensity difference was increased by 0, 5,10,15,20, 25, 30, 35 and 40 dB. This condition was very similar to the experiments conducted by Rand [4], the results reported by Rand [4] were just referred. Cutting [5] reported that there was no information loss or alteration in spectral fusion because the acoustic information of the opposite ear stimuli was simply restructured back into the original form from the component parts in a straight forward manner. For the third experiment a 24-item sequence was recorded to examine the robustness of spectral fusion when varying fundamental frequency. The first formant is held constant at 100 Hz while the second formant was varied as 100, 102, 120 and 180 Hz. Cutting [5] reported that the spectral fusion was immune to fundamental frequency variation. The listeners were able to identify the syllable clearly for all fundamental frequency variations. Cutting [5] attributed this to the fact that the fundamental frequency variation resulted in the two items being presented, but since the listeners had to choose one, they generally ignore the other one.

In order to substantiate this fact, Cutting [5] conducted the fourth part of the experiment in which the listeners were asked to report the number of sounds they heard. The stimuli used in the third experiment were used again in this experiment. The number of one-item responses dropped significantly as the fundamental frequency difference was increased. For the spectral fusion Cutting [5] reported a decrease of 58% for the change in fundamental frequency.

2.4 Effect of Fundamental Frequency

Darwin [6] study supported the results obtained by Cutting [5] . Darwin [6] investigated whether a common fundamental determined the grouping of the formants to form the correct syllables. Thus grouping of harmonics having the same fundamental would increase the intelligibility of the vowel, while the grouping of harmonics of different fundamentals deteriorated the intelligibility of the vowel.

The sounds used in his experiments were produced by filtering a train of clicks through parallel digital second-order filters. The resonant frequencies of the filter and the period and intensity of the click train entering each filter were under dynamic program control, but the bandwidths of the filters were fixed at 50, 60 and 100 Hz for formant 1, 2 and 3 respectively. Ten vowels were synthesized in which half of them had the same fundamental and in the other half the third formant had different fundamentals. The experiments were conducted with 10 normal hearing subjects. subjects had to report the clearest vowel they heard. Darwin [6] reported that the percentage correct for the same fundamental and different fundamental was almost at the same average percentage of 72% thereby supporting the report by Cutting [5].

Darwin [6] concluded that although varying the formant fundamental clearly increased the number of sounds that were reported, none of them had any consequences on the intelligibility of the vowel. Thus although the subjects heard multiple sounds, they were able to identify the vowel without any difficulty.

2.5 Dichotic Interleaved Presentation of Speech Signals

Analyses of dichotic speech perception led to various ways of implementing dichotic presentation. One of the most successful modes of presentation was the interleaved presentation wherein the speech signal was divided into number of frequency bands and all the odd bands were presented to one ear while the even bands were presented to the other ear.

Lunner *et al*, [7] compared the digital filter bank setup with the conventional analog filter used in the hearing aid. The filter banks were analyzed both for dichotic and diotic presentation. Lunner *et al*, [7] reported an improvement in the order of 2 dB in SNR for 50% correct recognition in dichotic condition when compared to the diotic condition.

The 8-band digital filter bank was an interpolated linear phase FIR filter with 40dB side lobes. The filter bank was implemented on a TI TMS320C25 Digital Signal Processor. 12 hearing impaired subjects are used for the experiments where the speech was presented diotically.

The experiment was repeated with the speech presented in dichotic mode. The experiment was conducted with 3 subjects. In dichotic mode, the odd bands were given to one ear and the even bands were presented to the other ear. Since the sensorineural hearing loss was commonly associated with reduced frequency selectivity, Lunner *et al*, [7] suggested that this interleaved mode of presentation might alleviate the problems causing the loss of frequency selectivity. Lunner *et al*, [7] also confirmed this suggestion through the result from the 3 subjects. With these results Lunner *et al*, [7] concluded that speech recognition in noise could possibly improved by splitting the total signal bandwidth with the odd bands being fed to one ear and the even bands to the other ear.

Pandey *et al*, [8] examined the effects of sensorineural hearing loss and extended the research of Lunner *et al*, [7] . Pandey *et al*, [8] reported that this sensorineural hearing loss might increase the spectral and temporal masking resulting in degraded speech perception. As mentioned in Lunner *et al*, [7] this could be attributed to the poor frequency selectivity. The increased spectral masking resulted in reduction of spectral contrast and also reduced the discrimination of consonant place feature. Pandey *et al*, [8] attempted to reduce the effect of the two maskings by using a pair of time varying comb filters for splitting the speech signals spectrally and temporally for dichotic presentation. The spectral splitting was implemented by simulating sensory cells corresponding to alternate bands of the basilar membrane. In temporal masking the ears get relaxed alternately for some time. Pandey *et al*, [8] combined both the splitting, resulting in a periodic relaxation from simulation for all the sensory calls of the basilar membrane.Simulation results showed that temporal and spectral splitting of frequency information could improve speech perception performance.

2.6 Summary

Summarizing the above researches, it can be seen that dichotic mode of presentation may help in improving speech recognition in noise and may also release the speech from masking. The spectral splitting of the speech has a significant effect in speech recognition. In order to investigate further the dichotic mode of speech presentation and the effects affecting dichotic listening, a set of experiments are performed first with normal listeners and then with the cochlear implant subjects. The experiments are presented in Chapter 3 and 4 respectively.

CHAPTER 3

DICHOTIC PRESENTATION FOR NORMAL LISTENERS

3.1 Chapter Outline

This chapter explains the experiments performed with normal hearing subjects. These experiments form the basis for the tests with cochlear implant subjects. The results of the implant subjects and the results are explained in the next chapter. Section 3.2 explains the details about the subjects participated in the tests. Section 3.3 lists the database used for the tests. Section 3.4 explains in great detail about the signal processing procedure behind simulating the cochlear implant type speech using MATLAB. Then section 3.5 feature the experimental procedure and the setup used. Section 3.6 displays the results. Section 3.7 concludes the chapter by analyzing the results displayed.

3.2 Subjects

Nine normal-hearing listeners (20 to 30 years of age) participated in this Experiment. All subjects are native speakers of American English. The subjects are paid for their participation. The subjects are undergraduate students from the University of Texas at Dallas.

3.3 Speech Materials

Subjects are tested on sentence, vowel and consonant recognition. The vowel test includes the syllables: ‘heed’, ‘hid’, ‘hayed’, ‘head’, ‘had’, ‘hod’, ‘hud’, ‘hood’, ‘hoed’, ‘whod’, ‘heard’ produced by male and female talkers. A total of 22 vowel tokens are used for testing, 11 produced by 7 male speakers and 11 produced by 6 female speakers [not all speakers produced all 11 vowels]. These tokens are a subset of the vowels used in [9] and are selected from a large vowel database [10] to represent the complete area of the vowel space. The consonant test uses sixteen consonants in /aCa/ context taken from the Iowa consonant test [1]. All /aCa/ syllables are produced by a male speaker.

The sentence test uses sentences from the HINT database [11]. Two lists, consisting of 20 sentences, are used for each condition, and different lists are used in all conditions.

3.4 Signal Processing

Speech material is first low-pass filtered using a sixth order elliptical filter with a cut-off frequency of 6000 Hz. Filtered speech is passed though a pre-emphasis filter (high-pass) with a cut-off frequency of 2000 Hz. This is followed by band-pass filtering into n (n=6, 8, 12) frequency bands using sixth-order Butterworth filters. The cut-off frequencies of the bandpass filters are given in Table 3.1. Logarithmic frequency spacing is used for n=6, and Mel frequency spacing (linear spacing up to 1000 Hz and logarithmic thereafter) is used for n=8, 12. The output of each channel is passed

	6 Channels		8 Channels		12 Channels	
Channels	F_L	F_H	F_L	F_H	F_L	F_H
1	300.0	487.2	261.9	526.6	191.6	356.8
2	487.2	791.1	526.6	857.3	356.8	549.5
3	791.1	1284.5	857.3	1270.4	549.5	774.3
4	1284.5	2085.8	1270.4	1786.5	774.3	1036.6
5	2085.8	3387.1	1786.5	2431.2	1036.6	1342.5
6	3387.1	5500.0	2431.2	3236.7	1342.5	1699.4
7			3236.7	4242.9	1699.4	2115.8
8			4242.9	5500.0	2115.8	2601.5
9					2601.5	3168.2
10					3168.2	3829.2
11					3829.2	4600.4
12					4600.4	5500.0

Table 3.1. The 3-dB cutoff frequencies (Hz) of the bandpass filters used in this study. FL and FH indicate the low and high cutoff frequencies respectively of the bandpass filters.

through a full-wave rectifier followed by a second order Butterworth low-pass filter with a cutoff frequency of 400 Hz to obtain the envelope of each channel output. Corresponding to each channel a sinusoid is generated with frequency set to the center frequency of the channel and with amplitude set to the root-mean-squared (rms) energy of the channel envelope, estimated every 4 ms. The phases of the sinusoids are estimated from the FFT of the speech segment as per [12]. No interpolation is done on the amplitudes or phases to smooth out any discontinuities across the 4-ms segments.

Two sets of sine waves are synthesized for dichotic presentation, one corresponding to the left ear and one corresponding to the right ear. The sine waves with frequencies corresponding to the left-ear channels are summed to produce the left-ear signal, and similarly, the sine waves corresponding to the right-ear channels are summed to produce the right-ear signal. In the condition, for instance, in which the low frequency information is presented to the left ear and the high-frequency information is presented to the right ear, the envelope amplitudes corresponding to channels $1 - n/2$ are used to synthesize the left-ear signal and the amplitudes corresponding to channels $n/2 + 1 - n$ are used to synthesize the right-ear signal. The levels of the two synthesized signals (left and right) are adjusted so that the sum of the two levels is equal to the rms value of the original speech segment. This is done by multiplying the left and right signals by the same energy normalization value. Hence, no imbalance is introduced, in terms of level differences or spectral tilt, between the left and right envelope amplitudes.

3.5 Procedure

The experiments are performed on a PC equipped with a Creative Labs SoundBlaster soundcard. Stimuli are played to the listeners either monaurally or dichotically through Sennheisers HD 250 Linear II circumaural headphones. For the vowel and consonant tests, a graphical user interface is used that enabled the subjects to indicate their response by clicking a button corresponding to the syllable played. For the sentence test, subjects are asked to write down the words they hear. The sentences are scored in terms of percent words identified correctly (all words were scored). During the practice session, the identity of the test syllables (vowels or consonants) and sentences is displayed on the screen.

At the beginning of each test the subjects are presented with a practice session in which the speech materials are processed through the same number of channels used in the test and presented monaurally in quiet and in +5 dB speech-shaped noise. For further practice, vowel and consonant tests are administered with feedback to each subject. Three repetitions are used in the feedback session. The practice session lasts approximately 2 hrs. After the practice and feedback sessions, the subjects are tested with the various dichotic and monaural conditions. The vowels and consonants are completely randomized and presented to the listeners six times. No feedback is provided during the test, and all the tests are done with speech embedded in +5 dB speech-shaped noise taken from the HINT database.

Two different dichotic conditions are considered. In the first condition, which we refer to as low-high dichotic condition, the low frequency information (consisting of half of the total number of channels) is presented to one ear, and the high-frequency information (consisting of the remaining half high-frequency channels) is presented to the other ear. In the second dichotic condition, which we call odd-even (or interleaved) dichotic condition, the odd-index frequency channels are presented to one ear, while the even-index channels are presented to the other ear. In the monaural condition, the signal is presented monaurally to either the left or the right ear (chosen randomly) of the subject. The order in which the conditions and number of channels is presented is partially counterbalanced between subjects to avoid order effects. In the vowel and consonant tests, there are 6 repetitions of each vowel and each consonant. The vowels and the consonants are completely randomized. A different set of 20-sentence lists is used for each condition.

Pilot data shows that the 12-channel condition in +5 dB S/N yielded performance close to ceiling. Hence, for the 12-channel condition, we performed additional listening experiments to assess whether subjects are indeed integrating the information from the two ears, or whether they are receiving sufficient information in each of the two ears alone. For comparison with the odd-even stimuli presented dichotically, two additional conditions are created. In the first condition, the odd-index channels are presented to the left ear alone, and in the second condition, the even-index channels are presented to the right ear alone. Similarly, for comparison with the low-high stimuli presented dichotically, the low-frequency channels (lower half number of channels) are presented to the left ear alone and the high-frequency channels are presented to the right ear alone. Sentences, vowels and consonants were processed though the one-ear conditions for the odd-even condition comparison. Vowels and consonants are also processed through the one-ear conditions for the low-high comparison. No sentences are processed for the low-high comparison since there are an insufficient number of unique sentences in the HINT database.

3.6 Results

The mean percent correct scores for sentence, vowel and consonant recognition is shown in Fig. 3.1 as a function of number of channels for different presentation modes.

3.6.1 Sentences

A two-way ANOVA with repeated measures, using spectral resolution (number of channels) and presentation mode (monaural, dichotic odd-even and dichotic low-

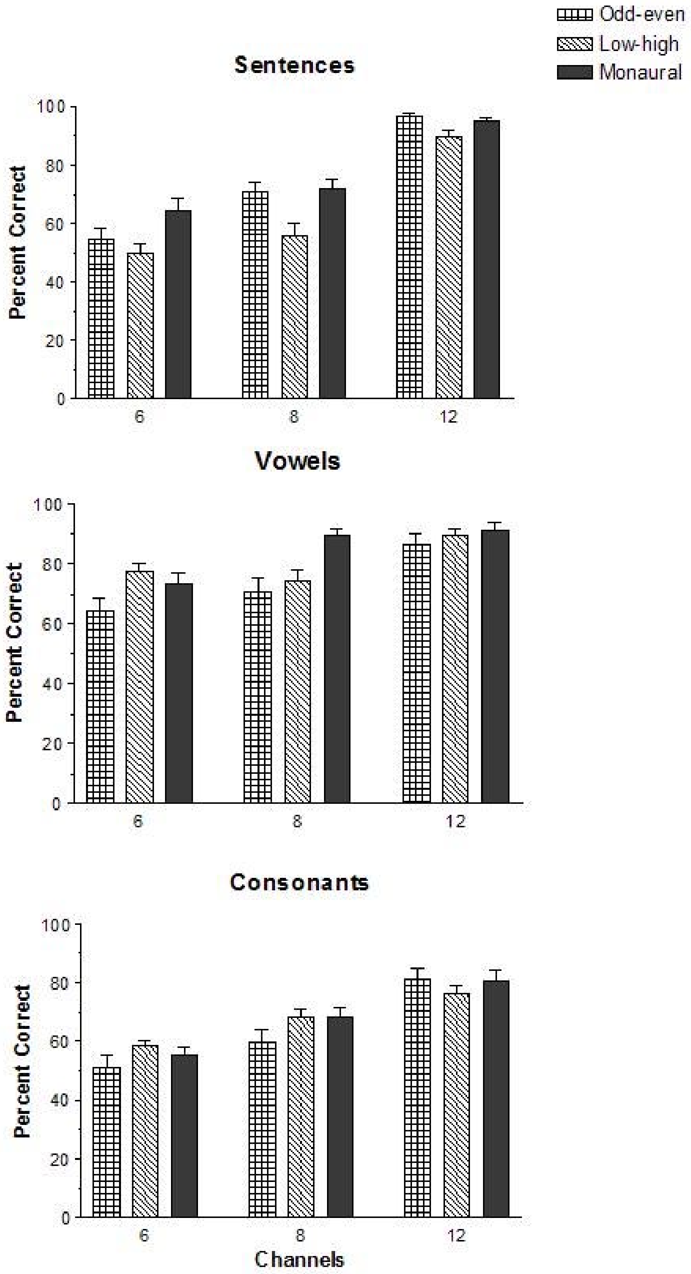


Figure 3.1. Mean identification of sentences (scored in percent words correct), vowels and consonants as a fraction of number of channels and presentation mode (monaural vs. dichotic). Error bars indicate standard errors of the mean.

high) as within-subject factors, shows a significant main effect of spectral resolution [F(2,16)=82.05, p<0.0005], a significant effect of presentation mode [F(2,16)=10.57, p=0.001], and a significant interaction [F(4,32)=3.36, p=0.02] between spectral resolution and presentation mode. Post-hoc tests according to Tukey (at alpha=0.05) shows that there is a significant (p<0.05) difference between the performance obtained with the two dichotic conditions for 12 and 8 channels, but not for 6 channels. There is also a significant difference (p<0.05) between the performance obtained with dichotic presentation (low-high) and monaural presentation for the 6 and 8 channel conditions, but not for the 12-channel condition.

Fig. 3.2 compares the dichotic performance (12 channels) obtained on sentence recognition with the performance obtained when the even channels are presented to the left ear alone, and the odd channels are presented to the right ear alone.

Post-hoc Fishers LSD tests showed that there is a significant difference (p<0.05) between the one-ear performance and the dichotic performance on sentence recognition, suggesting that subjects are able to integrate the information from the two ears. That is, the information presented in each ear alone is not sufficient to recognize sentences in +5 dB S/N with high (>90%) accuracy.

3.6.2 Vowels

A two-way ANOVA with repeated measures shows a significant main effect of spectral resolution [F (2, 16) =46.05, p<0.0005], a significant effect of presentation mode [F 2, 16) =4.79, p=0.023], and a significant interaction [F (4, 32) =7.16, p<0.0005] between spectral resolution and presentation mode on vowel recognition. Post-hoc

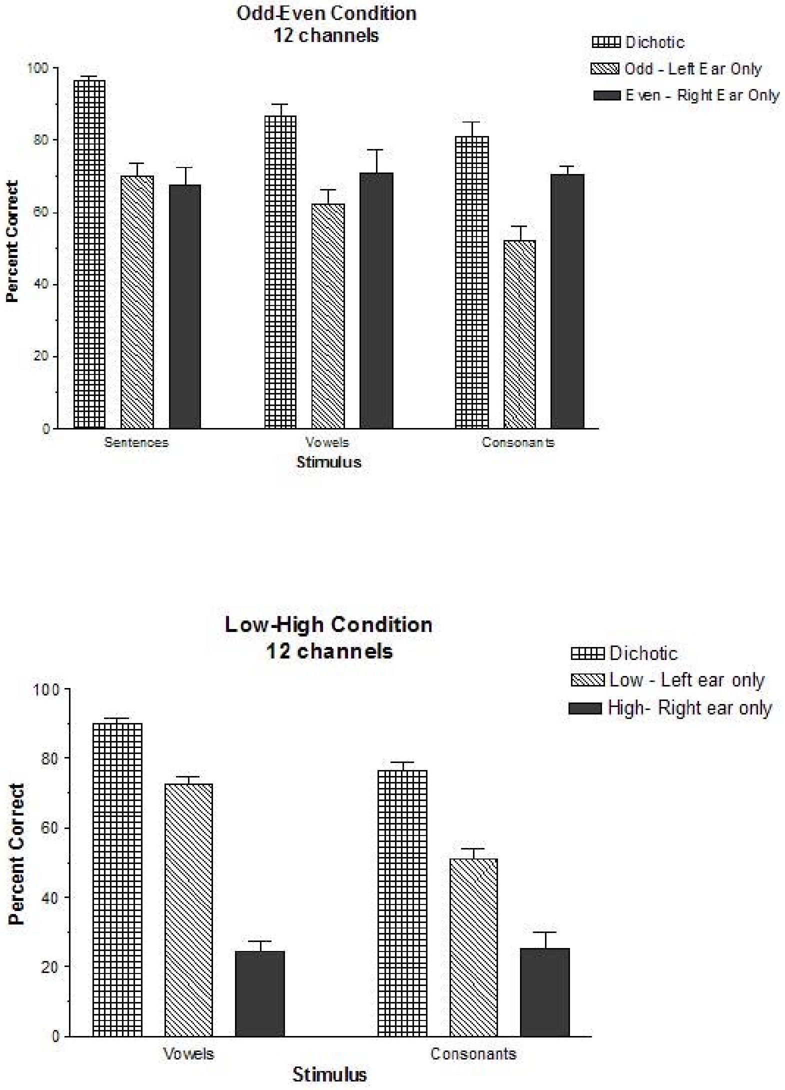


Figure 3.2. (Top panel) Mean speech recognition obtained when the odd- and even-index channels were presented dichotically (hatched bars), the odd-channels were presented to the left ear only (diagonally filled bars) and the even-channels were presented to the right ear only (dark bars). (Bottom panel). Mean vowel and consonant recognition obtained when the low- and high-frequency channels were presented dichotically (hatched bars), the low-frequency channels were presented to the left ear only (diagonally filled bars) and the high-frequency channels were presented to the right ear only (dark bars). Speech materials were processed through 12 channels. Error bars indicate standard errors of the mean.

tests according to Tukey shows that there is a significant ($p<0.05$) difference between each of the dichotic conditions and the monaural condition for 8 channels. There is also a significant difference between the two dichotic conditions for 6 channels.

Fig. 3.2 compares the performance obtained dichotically (for both conditions) with the performance obtained with the left and right ears only. Post-hoc Fishers LSD tests shows that there is a significant difference ($p<0.05$) between the one-ear performance and the dichotic performance on vowel recognition, suggesting that subjects are able to integrate the information from the two ears and obtain a vowel score higher than the score obtained with either ear alone.

3.6.3 Consonants

A two-way ANOVA with repeated measures shows a significant main effect of spectral resolution [F (2, 16) =46.05, $p<0.0005$], a non-significant effect of presentation mode [F (2, 16) =4.79, $p=0.34$], and a significant interaction [F (4, 32) =7.16, $p=0.002$] between spectral resolution and presentation mode on consonant recognition. As shown in Fig. 3.1, the presentation mode (dichotic vs. monaural) has no effect on consonant recognition.

Fig. 3.2 compares the consonant performance obtained dichotically against the performance obtained with either the left or right ears alone. Post-hoc Fishers LSD tests shows that there is a significant difference ($p<0.05$) between the one-ear performance and the dichotic performance on consonant recognition.

The consonant confusion matrices are also analyzed in terms of percent informa-

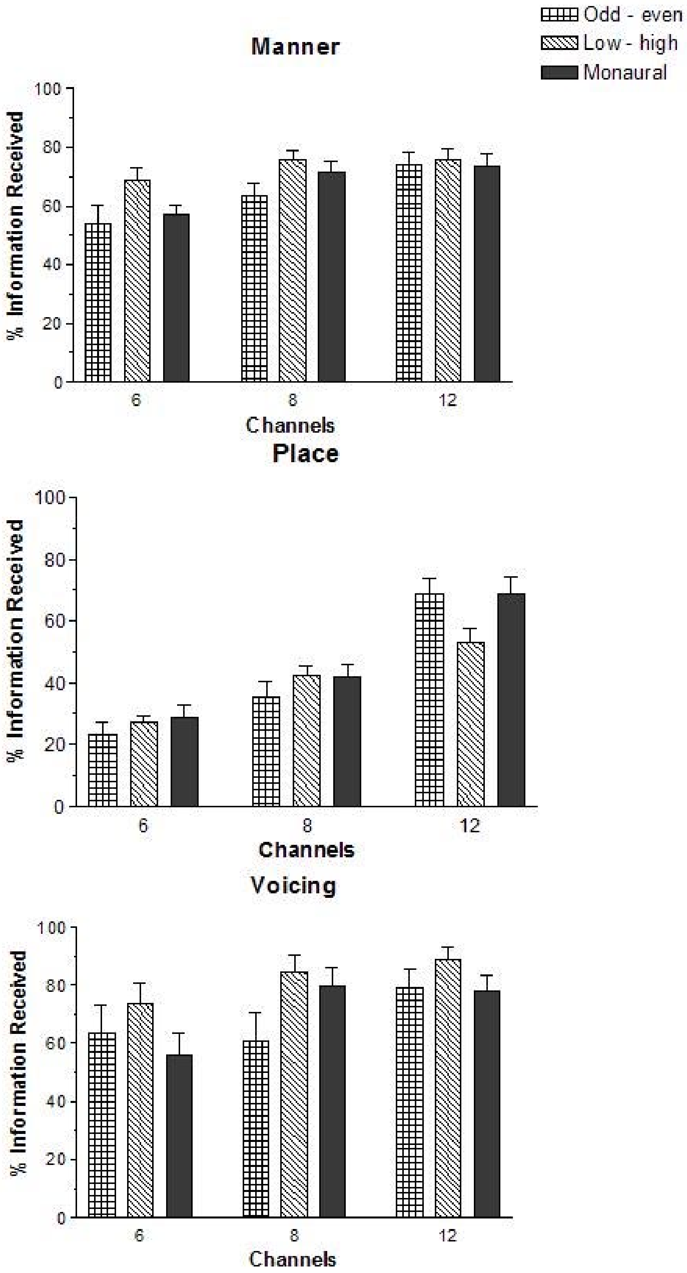


Figure 3.3. Mean identification of the consonant features manner, place and voicing as a function of number of channels and presentation mode (monaural vs. dichotic). Error bars indicate standard errors of the mean.

tion transmitted as per [13]. The feature analysis is shown in Fig. 3.3. A two-way ANOVA performed on each feature separately shows a marginally significant effect of presentation mode for manner ($p=0.04$) and voicing ($p=0.047$) and a non-significant effect for place ($p=0.31$). There is a significant interaction for place and voicing ($p<0.005$), but not for manner ($p=0.13$). There is significant effect ($p<0.005$) of spectral resolution for all three features.

3.7 Discussions

The above results indicate that the effect of presentation mode (dichotic vs. monaural) differed across speech materials and degree of spectral resolution. The recognition of sentences is affected the most. Recognition of vowels is also affected but to a lesser degree. As it is evident from the mean performance on sentences processed through 6 and 8 channels, subjects are not able to fuse sentences presented dichotically in the low-high condition with the same accuracy as when presented monaurally. Subjects are also not able to fuse vowels, processed through 8 channels and presented dichotically (in either low/high or even-odd conditions), with the same accuracy as when presented monaurally. It should be noted that there is large subject variability in performance (see Fig. 3.4) for vowels and sentences processed through 6 channels [see for instance subjects S1, S2 and S7 performance on sentence recognition, and subjects S4, S5 and S7 performance on vowel recognition]. In contrast, subjects are able to fuse vowels and sentences processed through 12 channels very accurately and more consistently. These results suggest that spectral resolution may have a significant effect on spectral fusion depending on the dichotic presentation mode (low/high vs.

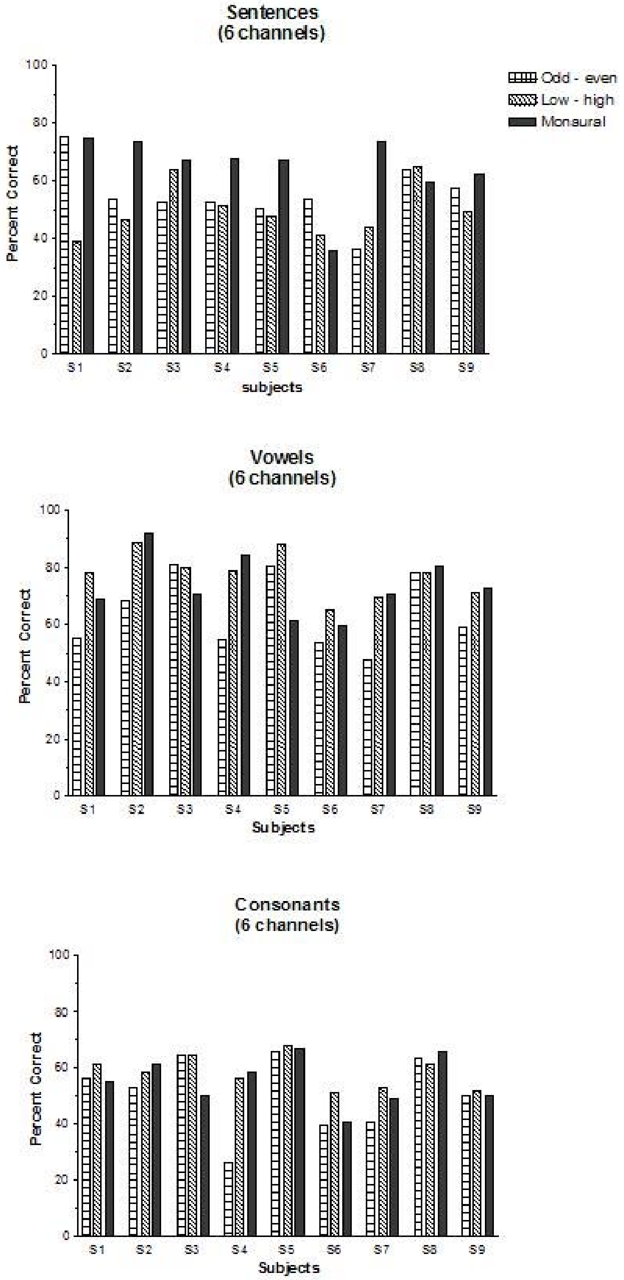


Figure 3.4. Individual subject performance for speech materials processed through 6 channels.

interleaved). No such effect is found for consonants. Consonants are fused accurately with both dichotic presentations regardless of the spectral resolution. The difference in effect on spectral fusion between consonants and the other speech materials suggests that vowels and sentences might be better (more sensitive) speech materials to use in studies of dichotic speech perception. This conclusion must be viewed with caution taking into account the fact the performance variability may be partially due to the variability in speech material used in this study, with the vowel material being more variable (produced my both male and female talkers, with multiple productions of each vowel) and the consonant material being less variable (produced by a single male talker with a single production of each consonant).

When interpreting the results on vowel recognition, two confounding factors need to be considered. First, the filter spacing is logarithmic for the 6-channel condition and Mel-like for the 8 and 12 channel conditions. We cannot exclude the possibility that a different outcome may have been obtained if different filter spacing is used, and this warrants further investigation. Second, the 8 and 12 channel conditions have a slightly larger (by about 38-108 Hz) overall bandwidth than the 6-channel condition.

Given that the roll-off of the 6th-order filters used is not too sharp, and the results from our previous study [14] indicating no significant difference on vowel recognition between different signal bandwidths, we do not believe that the additional bandwidth in the 8 and 12 channel conditions affected the outcome of this study.

No effect on spectral fusion is found in this study in the identification of consonants processed through a small number of channels. This outcome is consistent with the findings in the study reported by [15] with bilateral cochlear implant listeners. [15]

presented consonants in /aCa/ context to 2 bilateral implant users. The consonants are presented dichotically in the same two conditions used in our study, even-odd and low-high conditions. The bilateral subjects are fitted with 6- and 8-channel processors. Results shows a small advantage of dichotic stimulation, however the difference is not statistically significant. No experiments are done in [15] with other speech materials. Identification of consonants, in general, is known to be robust to extremely low spectral-resolution conditions ([16]; [17]) and even conditions in which large segments of the consonant spectra are missing ([18]; [19]). Dichotic identification of consonants does not seem to be an exception.

The fact that the listeners are not able to fuse with high accuracy sentences processed through 6 and 8 channels and presented in the low-high dichotic condition cannot be easily explained given that we does not manipulate in this study the relative onset time, the F0 or the relative intensity of the two signals presented to each ear.

One possible explanation is that the information presented in each ear is so impoverished, due to the high noise level and low spectral resolution (3 or 4 channels in each ear), that it is not perceived as being speech by the auditory system, hence it is not integrated centrally. This is based on the assumption that the ability to fuse a sound presented to one ear with another sound presented to the other ear must depend on recognizing that the two components come from the same speech utterance (and probably the same talker). In contrast, in the 12-channel condition the subjects has access to 6 channels of information in each ear, and we know from Fig. 3.2 that moderate levels of performance can be achieved in +5 dB S/N with only 6-channels presented to each ear. A similar outcome is also reported by [20] with HINT sentences presented dichotically in quiet with even channels fed in one ear and odd channels fed in the other ear. Sentences are processed through 6 channels in a manner similar to this study. Sentence scores obtained monaurally are significantly higher than the scores obtained dichotically.

In addition to observing a significant difference between the dichotic and the monaural conditions, a significant difference is also observed between the two dichotic conditions. The way spectral information is split and presented to the two ears is important for vowel and sentence recognition, but not for consonant recognition. For sentences processed through 8 and 12 channels, the mean scores obtained with the interleaved (odd-even) condition are significantly higher than the scores obtained with the low-high condition.

For vowels processed through 6 channels, the scores obtained with the low-high condition are significantly higher than the scores obtained with the interleaved condition. The higher scores obtained with the low-high condition in vowel recognition may be attributed to the fact that in this condition F1 information (low frequency information) is presented to one ear and F2 information (higher frequency information) is presented to the other ear. This condition must be easier to deal with compared to the more challenging condition (interleaved) in which pieces of F1 and F2 information are distributed between the two ears. Subjects did not have difficulty piecing together the F1/F2 information, however, when the vowels are processed through 8 and 12 channels. The above explanation can not be easily extended to sentences, because the situation with sentences differs from that of vowels, in that listeners are relying on other cues, besides F1/F2 information, for word recognition.

One possible explanation for the higher scores obtained in sentence recognition with the interleaved condition is that it provides greater dichotic release from masking (a phenomenon first reported in [4]) compared to the low-high condition. Rand [4] presented the F1 of CV syllables to one ear, and the F2 attenuated by 40 dB to the other ear, and observed that subjects are able to identify the consonants accurately despite the large relative intensity differences between the two formants. However, when he attenuated the upper formants by 30 dB and presented the stimuli to both ears (i.e., diotically), the listeners are unable to identify the consonants accurately. Rand [4] attributed the advantage of dichotic presentation to release from spectral masking.

Others (e.g., [7], [8], [21], [22]) have attempted to exploit the dichotic release from masking in bilateral hearing aids and advocated the use of dichotic presentation as a means of compensating for the poor frequency selectivity of listeners with sensorineural hearing loss. Lunner *et al*, [7] fitted three hearing-impaired listeners with 8-channel hearing aids and presented sentences dichotically, with the odd-number channels fed to one ear and the even-number channels fed to the other ear. An improvement of 2 dB in speech reception threshold (SRT) is found compared to the condition in which the sentences are presented binaurally. Franklin [22] investigates the effect of presenting a low-frequency band (240-480 Hz) and a high-frequency band (1020-2040 Hz) on consonant recognition in six hearing impaired listeners with moderate to severe sensorineural hearing loss. The scores obtained when the low and high frequency bands are presented to opposite ears are significantly higher than the scores obtained when the two bands are presented to the same ear. Unlike the above two studies, a few other studies (e.g., [21], [23]) finds no benefit of dichotic presentation with hearing-impaired listeners tested on synthetic /b d g/ identification [23] or vowel/consonant identification [21].

In the present study, speech is synthesized as a sum of a small number (6-12) of sine waves. Sine wave speech is advocated in our earlier work [17] as an alternative to noise-band simulations, [16] acoustic model for cochlear implants. This acoustic model is not meant by any means to mimic the percept elicited by electrical stimulation, but rather to serve as a tool in assessing CI listener's performance in the absence of confounding factors (e.g., neuron survival, electrode insertion depth, etc.) associated with cochlear implants. We previously demonstrated on tests of reduced intensity resolution [24], reduced spectral contrast [25] and number of channels of stimulation [26] that the performance of normal-hearing subjects, when listening to speech processed through the sine wave acoustic model, is similar to that of, at least, the better performing patients with cochlear implants. In that context, the results of the present study might provide valuable insights to bilateral cochlear implants. The performance obtained with the interleaved dichotic condition is not significantly different from the performance obtained monaurally when the speech materials were processed through 12 channels.

For bilateral patients receiving a large number of channels (>12), dichotic stimulation might therefore provide an advantage over unilateral stimulation, in that it can potentially reduce possible channel interactions since the electrodes can be stimulated in an interleaved fashion across the ears without sacrificing performance. For patients receiving only a small number of channels of information, dichotic (electric) stimulation might not produce the same level of performance as unilateral (monaural) stimulation, at least for sentences presented in the low-high dichotic condition. The large variability in performance among subjects should be noted however (Fig. 3.4). Lastly, of the two methods that can be used to present spectral information dichotically, the interleaved method is recommended since it consistently outperformed the low-high method in sentence recognition.

CHAPTER 4

DICHOTIC SPEECH RECOGNITION BY BILATERAL COCHLEAR IMPLANTS LISTENERS

4.1 Chapter Outline

This chapter outlines the experimental setup used for the experiments with cochlear implant subjects. Section 4.2 gives a brief introduction to the Spear 3 processor used in these experiments, and Section 4.3 tabulates the profile of the subjects participated in the tests. Section 4.4 describes the speech materials used, Section 4.5 explains the experimental setup, and Section 4.6 explains the procedure followed for the tests. Section 4.7 summarizes the results. Section 4.8 explains the effect of beamforming on the intelligibility of Cochlear Implant listeners. Section 4.9 describes the beamforming experiments. Section 4.10 summarizes the results. Section 4.11 analyzes the results.

4.2 SPEAR3 Research Processor

4.2.1 Introduction

The Speech Processor for Electrical and Acoustic Research (SPEAR3) is a speech processor designed primarily for cochlear implant and hearing aid research. With 2-channel audio I/O and twin implant drivers, the SPEAR3 embodies many of the features required for bilateral implant or hearing aid research in a portable device.

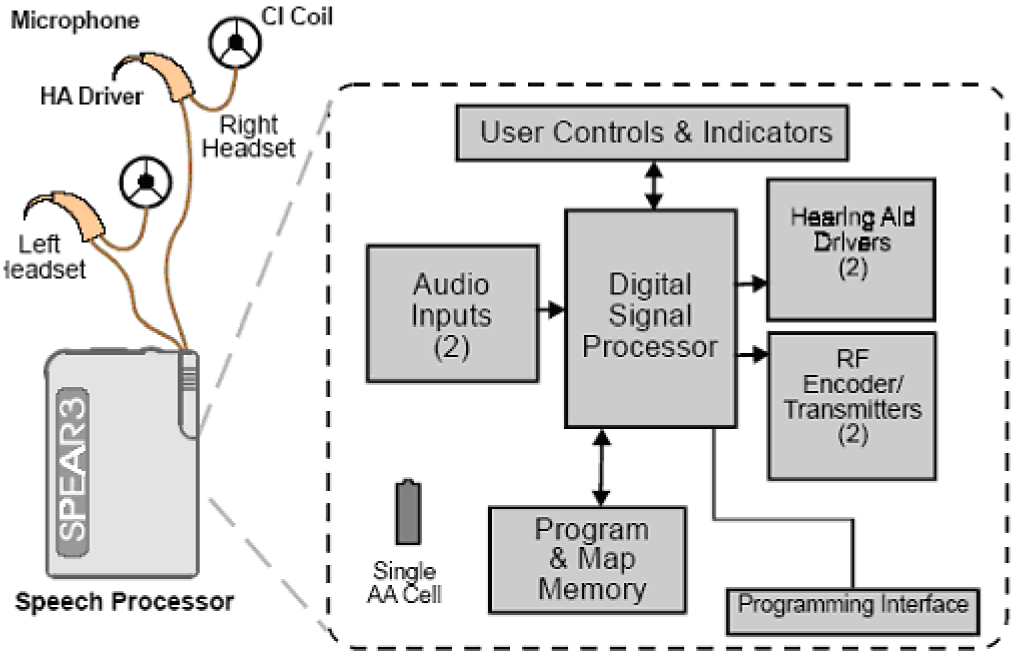


Figure 4.1. SPEAR3 Functional Diagram

4.2.2 Functional Description

The speech processor consists of a number of blocks as shown in the Fig. 4.1. Each block is described in the sub sections.

4.2.2.1 Audio Section Input

The SPEAR3 processor accepts low-level audio from two headset microphones. Alternatively an external mono or stereo microphone can be plugged into a standard 3.5mm socket. A stereo CODEC converts the audio input to linear 16-bit words for processing by the DSP. The CODEC uses 16-bit Sigma-Delta conversion after performing anti-alias filtering. Data from the DSP are converted by the CODEC back to an analog audio signal and amplified by a power amplifier capable of driving two low impedance hearing aid drivers. It may also drive stereo headphones to aid software debugging.

4.2.2.2 DSP

The Motorola DSP56309 24-bit digital signal processor is at the core of the speech processor. It performs all signal processing tasks, such as filtering, spectral analysis and coding functions. The DSP also performs house keeping functions such as monitoring front panel controls, measuring battery states, and communicating with a host PC during configuration.

A Flash RAM provides non-volatile storage for program storage and data such as patient maps. This device can also be used for storing user settings when the processor is turned off. A serial data link is used to program the speech processor from a PC and perform limited diagnostic tasks. A simple interface (SPS) connects the SPEAR3 to a standard PC serial port. The SPEAR3's DSP also has a OnCE port that can be connected to a host system for more advanced program development and debugging using standard DS P56300 development tools.

4.2.2.3 Implant Transmitter

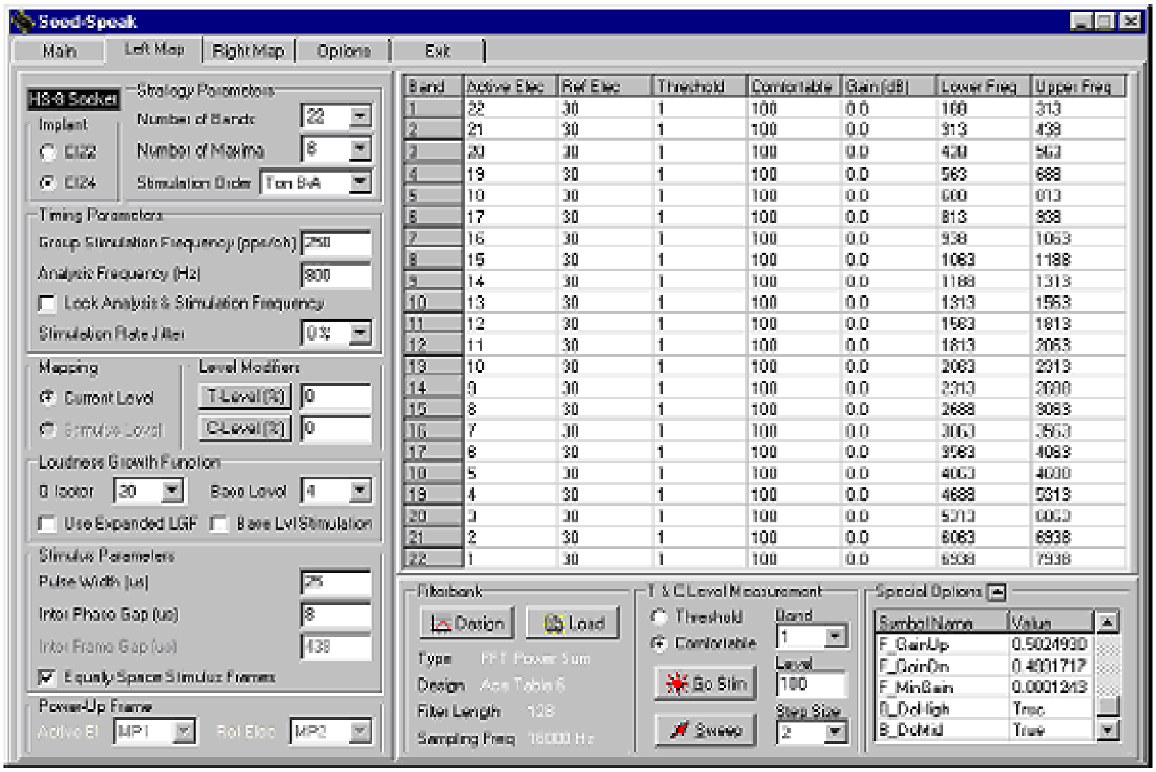
The SPEAR3 contain two identical but independent transmitter sections that can be used to drive Cochlear's 22 and 24 channel cochlear implant devices. These sections receive stimulus data from the DSP and encode the data into a serial data stream which modulates a radio frequency driver which in turn drives the coil of the implant headset.

The SPEAR3 may be used with a single implant but is also capable of being used by bi-laterally implanted users. A combination of 22 and 24 channel implants can be accommodated. Since the transmitter is independent of the Audio output section, it is also possible to use the SPEAR3 with combinations of implants and hearing aids, in either “combionic” configurations, or residual hearing configurations.

4.2.3 SPEAR3 Speech Processing System

The Speech Processing System provides a means for programming the SPEAR3 processor for speech coding research with cochlear implants. The system comprises two components: (1) a core speech coding program for the SPEAR3 processor known as CSPEAK (Configurable SPEAK), and (2) a MS-Windows program known as Seed-Speak, which configure the parameters of CSPEAK, and programs the SPEAR3 processor. The system supports unilateral and bilateral implant configurations. Many program and client parameters, such as rate of stimulation, analysis (update) rate, number of bands (channels), and number of maxima, loudness growth mapping parameters, and subjective threshold and comfortable stimulation levels can be configured by the system.

The core speech coding program (CSPEAK) is a Configurable version of the SPEAK strategy that can be configured to emulate the SPEAK, ACE or CIS strategies. The Seed-Speak (MS-Window) application is used to configure parameters of the core program, such as those described above, as well as to provide a mechanism for maintaining client (patient) maps and to load these maps into the SPEAR3 processor. Seed-Speak can also be used to configure the filter bank used by the core program.



Seed-Speak

Main | Left Map | Right Map | Options | Exit

HS-8 Socket

Implant
○ CI22
● CI24

Strategy Parameters
Number of Bands: 22
Number of Maxima: 8
Stimulation Order: Ton B-A

Timing Parameters
Group Stimulation Frequency (pps/ch): 250
Analysis Frequency (Hz): 800
☐ Lock Analysis & Stimulation Frequency
Stimulation Rate Jitter: 0%

Mapping
● Current Level
○ Stimulus Level

Level Modifiers
T-Level(%): 0
C-Level(%): 0

Loudness Growth Function
Q factor: 20
Base Level: 4
☐ Use Expanded LGF
☐ Base Lvl Stimulation

Stimulus Parameters
Pulse Width (us): 25
Inter Phase Gap (us): 8
Inter Frame Gap (us): 438
☑ Equally Space Stimulus Frames

Power-Up Frame
Active El: MP1
Ref Elec: MP2

Band	Active Elec	Ref Elec	Threshold	Comfortable	Gain (dB)	Lower Freq	Upper Freq
1	22	30	1	100	0.0	188	313
2	21	30	1	100	0.0	313	438
3	20	30	1	100	0.0	438	563
4	19	30	1	100	0.0	563	688
5	18	30	1	100	0.0	688	813
6	17	30	1	100	0.0	813	938
7	16	30	1	100	0.0	938	1063
8	15	30	1	100	0.0	1063	1188
9	14	30	1	100	0.0	1188	1313
10	13	30	1	100	0.0	1313	1563
11	12	30	1	100	0.0	1563	1813
12	11	30	1	100	0.0	1813	2063
13	10	30	1	100	0.0	2063	2313
14	9	30	1	100	0.0	2313	2688
15	8	30	1	100	0.0	2688	3063
16	7	30	1	100	0.0	3063	3563
17	6	30	1	100	0.0	3563	4063
18	5	30	1	100	0.0	4063	4688
19	4	30	1	100	0.0	4688	5313
20	3	30	1	100	0.0	5313	6063
21	2	30	1	100	0.0	6063	6938
22	1	30	1	100	0.0	6938	7938

Filterbank
Design | Load
Type: FFT Power Sum
Design: Ace Table 6
Filter Length: 128
Sampling Freq: 16000 Hz

T & C Level Measurement
○ Threshold
● Comfortable
Band: 1
Level: 100
Step Size: 2
Go Stim | Sweep

Special Options

Symbol Name	Value
F_GainUp	0.5024930
F_GainDn	0.4001717
F_MinGain	0.0001243
B_DcHigh	True
B_DcMid	True

Figure 4.2. Seed Speak Client Map Settings Window

The filter bank can be matched to that used by the ACE or SPEAK (SPrint) speech coding strategies or designed to some other specification. In addition, Seed-Speak provides some psychophysics functions for measurement of subjective threshold and comfortable loudness levels, sweeping of electrodes, electrode-pitch estimation and loudness balancing. Fig. 4.2 shows the screen shot of Seed-Speak application.

4.3 Subjects

Four bilateral implant subjects who use the Nucleus24 device were tested. The biographical data of these subjects are shown in table 4.1.

Subject	Age	Etiology of Deafness	Years Implanted		Duration of Deafness
			Left	Right	
S1	38	Unknown	2002	2002	2 years
S2	81	Progressive Hearing Loss	2001	2001	30 years
S3	44	Auto Immune Disease	5/2001	5/2001	5 months
S4	34	Possible cause-Diabetics	10/2002	12/2002	10 years

Table 4.1. Biographical data of subjects.

4.4 Speech Materials

HINT database sentences [11] are used for the testing in quiet and the TIMIT database sentences [27] are used for sentence recognition in noise. The sentences are corrupted with speech shaped noise at +10dB. Thirty sentences were used for each condition. The TIMIT sentence list were normed in our lab for equal intelligibility.

4.5 Experimental Setup

All the experiments are carried out using the SPEAR3 bilateral cochlear implant mentioned in the section 4.2. The CSPEAK program was used for programming the experimental MAPs. The pulse rate, pulse width and interphase gap were set to the values used in the subjects daily strategy. To ensure that a fixed number of electrodes were stimulated in each cycle, we fitted subjects with a 12-channels CIS. The sentences were played through a MATLAB program and presented to the implant listners at a comfortable level via the SPEAR3's auxiliary input jack.

4.6 Procedure

Nine sets of experiments were performed. The experiments included two dichotic conditions and a baseline condition for comparison. In order to examine spectral fusion, two additional experiments were performed. In this set of experiments, the speech material is presented to one ear only. All the above experiments are conducted both in quiet as well as in the presence of speech shaped noise at +10dB SNR.

In the first dichotic conditions, which we refer to as low-high dichotic condition, the low frequency information (consisting of half of the total number of channels) is presented to one ear, and the high-frequency information (consisting of the remaining half high-frequency channels) is presented to the other ear. In the second dichotic condition, which we call odd-even (or interleaved) dichotic condition, the odd-index frequency channels are presented to one ear, while the even-index channels are presented to the other ear. In the monaural condition, the signal is presented monaurally to both the left and the right ears of the subject. The order in which the conditions are presented is partially counterbalanced between subjects to avoid order effects. As explained previously, additional conditions are conducted to examine spectral fusion. For comparison with the odd-even stimuli presented dichotically, the odd-index channels are presented to the left ear alone, and in the second condition, the even-index channels are presented to the right ear alone. Similarly, for comparison with the low-high stimuli presented dichotically, the low-frequency channels (lower half number of channels) are presented to the left ear alone and the high-frequency channels are presented to the right ear alone. For comparison with the baseline condition, all 12

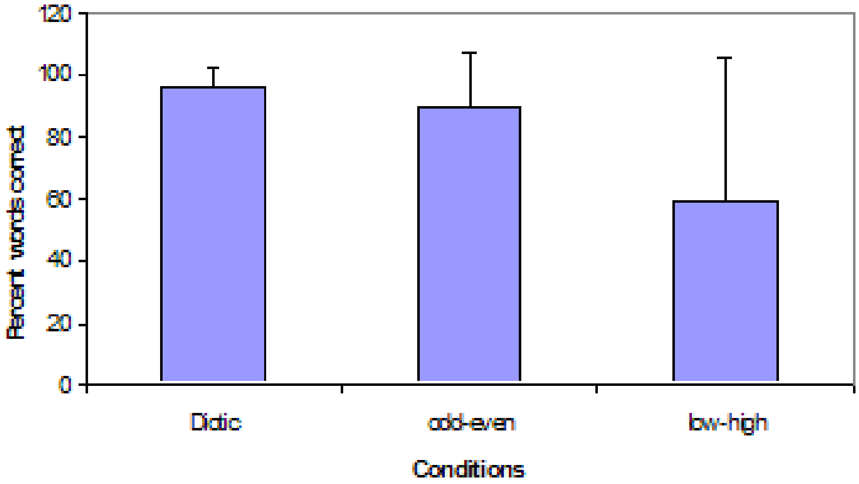


Figure 4.3. Mean scores for the diotic and the two dichotic condition in quiet.

channels are presented to the left ear alone or all the 12 channels are presented to the right ear alone.

4.7 Results

4.7.1 Results in Quiet

Fig. 4.3 compares the results obtained in the three conditions namely the monaural and the two dichotic conditions. It is clearly seen that there is no large difference between the scores obtained with the diotic condition and the scores obtained with the two dichotic conditions. This suggests that the subjects were able to fuse the information presented to both ears.

Fig. 4.4 and Fig. 4.5 further demonstrate spectral fusion. Fig. 4.4 shows performance obtained when the odd-channels or even-channels were presented to each ear

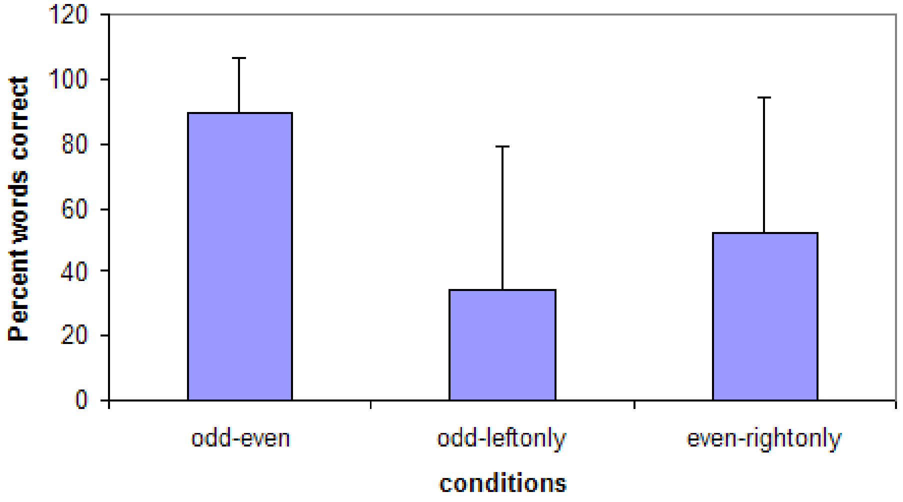


Figure 4.4. Mean speech recognition obtained when the odd- and even-index channels are presented dichotically, the odd-channels are presented to the left ear only and the even-channels are presented to the right ear only. All the tests are conducted in quiet.

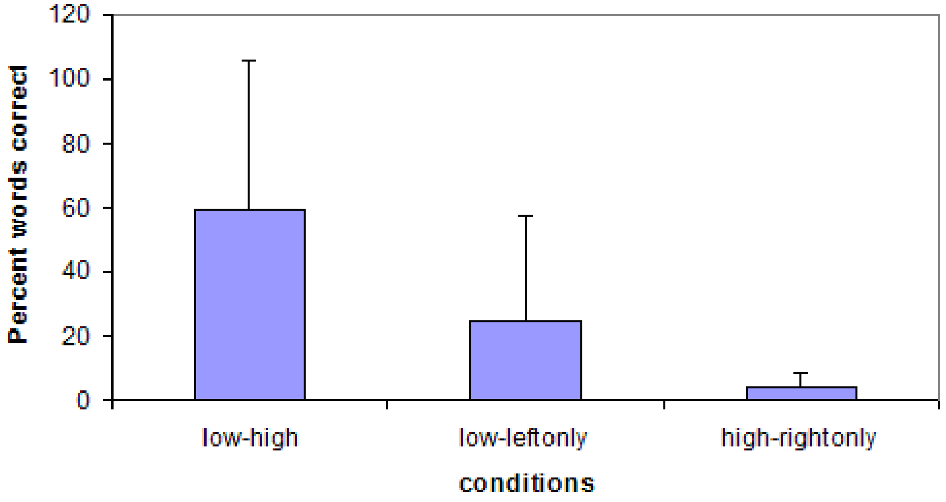


Figure 4.5. Mean speech recognition obtained when the low- and high-index channels are presented dichotically, the low frequency-channels are presented to the left ear only and the high frequency-channels are presented to the right ear only. All the tests are conducted in quiet.

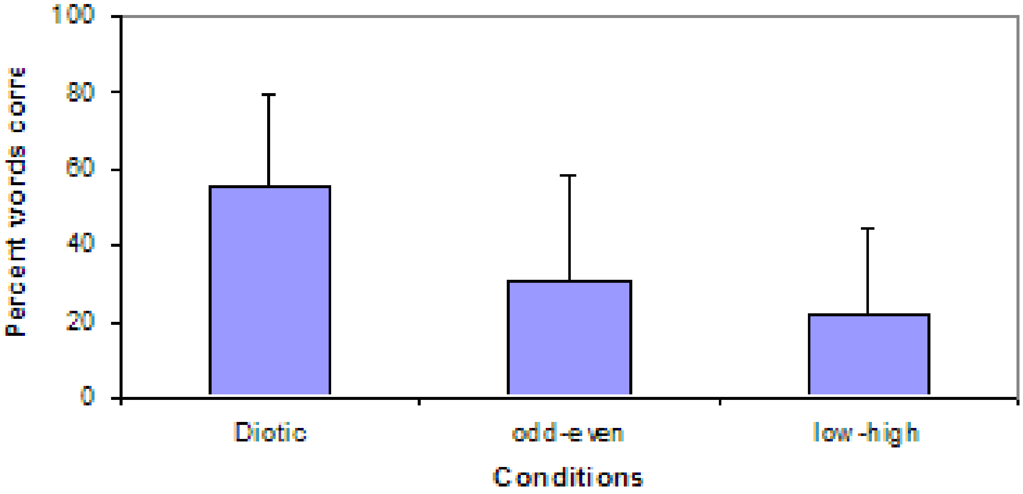


Figure 4.6. Mean score for the diotic and two dichotic conditions in noise.

alone. Fig. 4.5 shows performance obtained when the low-frequency channels or high-frequency chan nels were presented to each ear alone. As can be seen, performance obtained dichotically is significantly higher than performance obtained with either ear alone, suggesting that subjects were able to spectrally fuse the information presented dichotically.

4.7.2 Results in Noise

Fig. 4.6 compares the results for the two dichotic conditions in noise with the baseline condition (diotic condition). In noise, subjects performed poorly in the dichotic condition compared to the diotic condition ($p<0.005$). Fig. 4.7 and Fig. 4.8 provide a closer look on the fusion of speech signals presented dichotically.

Comparing Fig. 4.6 and Fig. 4.7 with Fig. 4.8 we see that noise disrupts the fusion of the information presented dichotically. Moreover the large error bars in

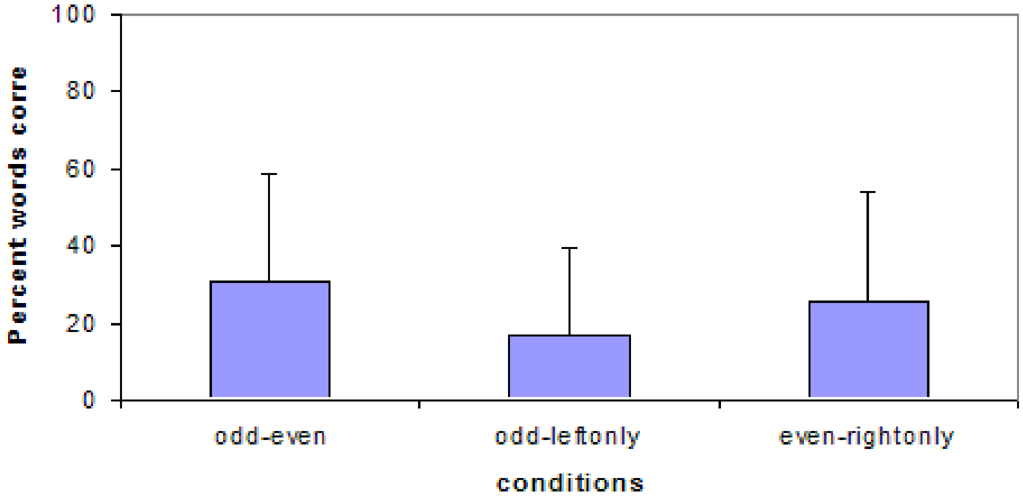


Figure 4.7. Mean speech recognition obtained when the odd- and even-index channels are presented dichotically, the odd-channels are presented to the left ear only and the even-channels are presented to the right ear only. All the tests are conducted in noise with SNR of +10dB.

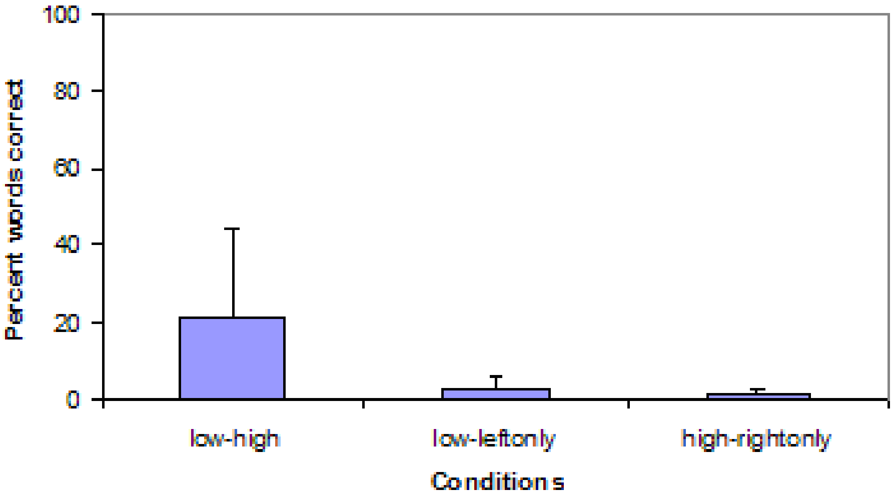


Figure 4.8. Mean speech recognition obtained when the low- and high-index channels are presented dichotically, the low frequency-channels are presented to the left ear only and the high frequency-channels are presented to the right ear only. All the tests are conducted in noise with SNR of +10dB.

these graphs suggests that the performance in the dichotic condition is largely subject dependent. Performance is greatly dependent on the ability of the subjects to retrieve the information masked by the noise and then fuse it.

4.8 Effects of Beamforming on Intelligibility

4.8.1 Introduction

Beamforming techniques have been applied extensively to hearing aids [28]. Greenberg *et al*, [29] investigated the effects of adaptive beamforming methods for the bilateral hearing aids. They concluded that depending upon the acoustic environment, the beamforming improves the intelligibility of the bilateral hearing aids. Motivated by these findings, the same beamforming technique is investigated for the bilateral cochlear implants.

4.8.2 Head Related Transfer Function

Consider the situation shown in Fig 4.9 in which the source of the speech reaching a person comes from his right side at an angle of θ degrees from the center. The speech signal from the source reaching the ear on the right does not undergo significant change, but the signal reaching the other ear undergoes large change mainly due the head blocking the direct waves reaching the ear. The blocking of the waves by the head, is termed as the *headshadow* effect. Many efforts were made to model the head shadow effect. One approach is to consider the head shadow effect as a linear filtering system. This in turn helps us model the head shadow effect by a linear FIR tap filter.

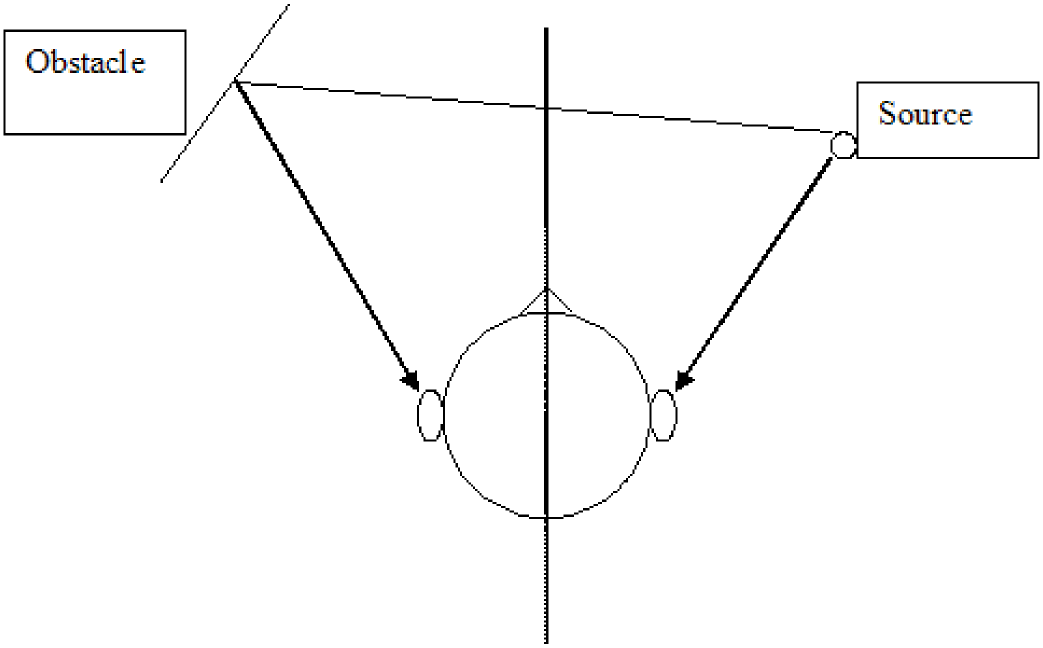


Figure 4.9. Head Shadow Effect.

Then the problems in modeling the system narrows down to determining the impulse response of the FIR filter.

Gardner *et al*, [30] developed a database of speech signals reaching the two ears from different elevations and azimuths (angles). The Knowles Electronic Model for Audio Research (KEMAR) dummy was used for their work. The microphones were mounted behind the ear of the dummy and placed at the center of the sound source. The sound source position was varied and for each position, a pair of speech signal was recorded by the microphones. This work assumed that the torso and pinnae effects were not present.

Since the head is modeled as a FIR filter, the signal that is received at the ear on the side of the source can be considered as the input to the filters, while the speech received by the other microphone can be considered as the output of the filter. So the impulse response can be obtained from these two signals as according to :

$$H(\omega) = \frac{Y1(\omega)}{Y2(\omega)} \tag{4.1}$$

where $H(\omega)$, is the frequency response of the filter and $Y1(\omega),\ Y2(\omega)$ are the spectra of the input and output signals respectively. From 4.1 we can easily obtain the impulse response of the filters for different angles. This impulse response can be used to calculate the head shadowed signal for the speech samples considered.

4.9 Beamforming Experiments

4.9.1 Generating the Noisy Binaural Speech

The binaural speech signals generated for the beamforming experiments consist of signals from two sources. The first source contains the speech, and the second source contains the noise which is multi-talker babble. The speech source is considered to be in the front while the noise source is at 45 degrees to the right from the center. The speech signal reaching the two ears is the same (assuming anechoic environment) since the source is at the center, while the noise source is altered after reaching the left ear. So the noise is passed through the HRTF to produce the left channel noise and then added to the speech samples from the source to generate the binaural noisy speech. The noisy speech has an SNR of 5dB.

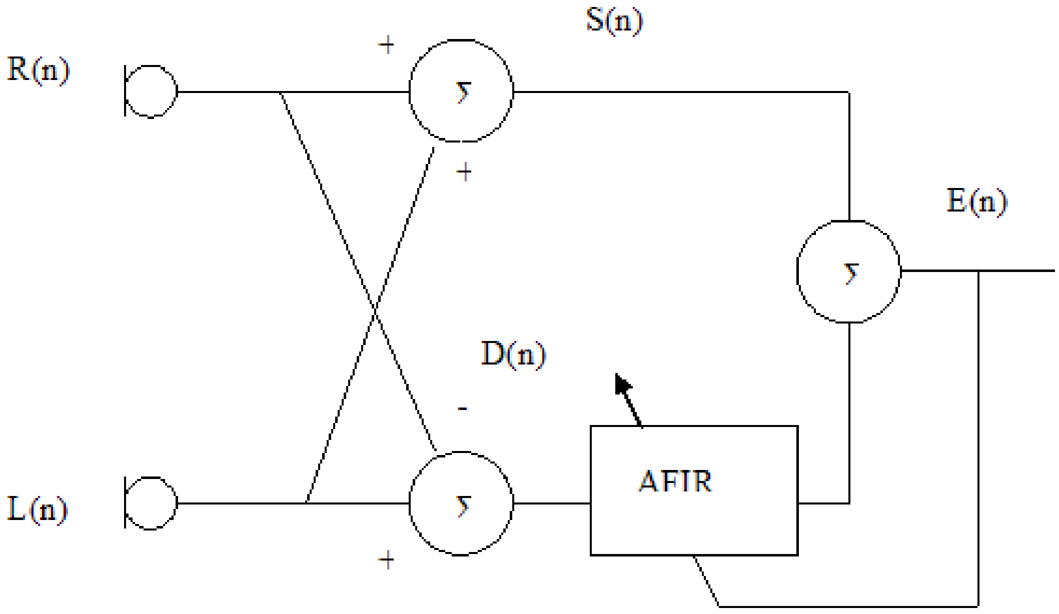


Figure 4.10. Beamforming using Jim Griffiths algorithm.

4.9.2 Beamforming

4.9.2.1 Griffith-Jim Algorithm

After generating the binaural speech samples as explained above, the speech samples are used as input to the beamforming. VanHoesel *etal*, [28] used the Griffith-Jim algorithm [31] for beamforming. The Griffith-Jim algorithm is the simplest form of beamforming, and is shown in Fig 4.10. The two inputs are the left and right channel signals generated above. The inputs are added and subtracted and the sum and difference signals are fed into the block as shown in Fig 4.10. The difference signal is passed through an LMS adaptive FIR filter (AFIR) and the output of this filter is subtracted from the sum signal to obtain the enhanced signal. This difference is used to adapt the filter coefficients continously. Let R(n) be the right ear channel and L(n) be the left ear channel. These two signals are added and subtracted to yield the S(n) and D(n) signals respectively as shown Fig 4.10. The D(n) is then passed through the LMS filter and the output is subtracted from S(n) to yield the error E(n). The E(n) signal is the desired output and also used to update the coefficient of the filter. The weights are updated according to the following function:

$$w(n+1) = w(n) + 2\mu E(n) \tag{4.2}$$

where (μ =0.09) is the step size of the LMS filter and $w(n+1)$, $w(n)$ are the updated and current coefficients of the LMS filter respectively.

The filter is updated using the equation 4.2 whenever equation 4.3 is satisfied. [28]:

$$Thr_S(n) - 2Thr_D(n) > 0 \tag{4.3}$$

where

$$Thr_S(n) = \alpha S(n) + (1-\alpha)S^2(n) \tag{4.4}$$

$$Thr_D(n) = \alpha D(n) + (1-\alpha)D^2(n) \tag{4.5}$$

where (α = 0.75) is the forgetting factor. The error signal E(n) is the enhanced output signal. The number of taps of the adaptive FIR filter affects the output of the system. After analyzing various tap sizes we considered using 256 tap filters.

4.9.2.2 Procedure

Three conditions are considered in these experiments. In the first condition, we used as input the noisy binaural speech files generated by the HRTF procedure explained

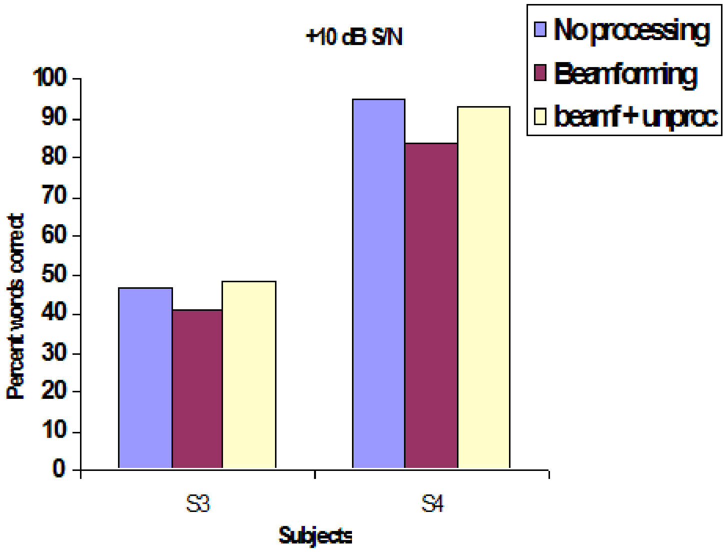


Figure 4.11. Individual scores obtained when unprocessed speech presented binaurally, beamformed output, beamformed output in one ear and the unprocessed speech in the other ear.

above. This condition was considered to be the baseline condition. In the second condition, noisy binaural files were input to the beamforming algorithm to produce the enhanced signals. The speech received in both ears is the same and since the signal was an enhanced version of the noisy signal, the signal to noise ratio is improved.

In the third condition, the enhanced speech signal is provided to the better ear, while the raw noisy speech signal is fed into the other ear. This condition is meant to examine whether the noise in one ear can mask the residual present in the other ear. In order to reduce the masking effect of the noise, the power of the noisy speech signals is halved. The order of these conditions was counter balanced across subjects.

4.10 Results from Beamforming

Fig 4.11 depicts the trend obtained from the beamforming tests. This experiment was done only with subjects 3 and 4. So it will be highly unreliable to conclude from the graph. But it is clearly seen that the condition in which the beamformed speech in one ear and unprocessed in the other performs slightly better than the condition in which beamformed speech is given to both ears.

4.11 Discussion

From the results above, we conclude that in quite, the dichotic performance is comparable to the diotic performance. In noise, however dichotic performance was worse than the diotic performance. Even normal hearing subjects had difficulty fusing information in noise when presented with a small number of channels. We therefore conclude that the inability of bilateral cochlear implant users to fuse information in noise must be due to the small number of channels received in each ear. The interleaving (odd-even) dichotic condition generally performed better than the low-high condition. This supports the inference that interleaving the spectral information between the ears is more effective than dividing the information into low and high frequency content and presenting it to the two ears.

The experiments with beamforming clearly show that providing the beamformed signal to both ears decreases the intelligibility compared to the case when beamformed signal is presented to one ear and the unprocessed speech in the other ear. We believe that this is due to the fact that the former case disrupts the ILD cues present in the signal, as it presents the identical , albeit enhanced, signal to both ears. More subjects are needed however to make any reliable conclusions about the benefit of beamforming.

CHAPTER 5

SUMMARY AND CONCLUSIONS

This study focused on analyzing the dichotic presentation of speech signals to binaural cochlear implant users. Past research concentrated on different dichotic modes and the factors affecting dichotic presentation. This thesis extended the existing findings and investigated the effects of spectral resolution on dichotic presentation.

This thesis starts with testing the spectral resolution effect on dichotic listening in normal hearing listeners. The test samples are processed through a cochlear implant simulation and presented to the listeners through headphone. The study is then extended to examine the effect of spectral resolution on bilateral cochlear implant subjects. Speech is presented in two different dichotic modes which are termed as low-high and odd-even. In the low-high condition, the bands corresponding to the low frequencies are presented to one ear while the bands corresponding to the high frequencies are given to the other ear. In case of odd-even the odd numbered bands are given to one ear and the even numbered bands are given to the other ear. The major conclusions from these studies were:

- There was a significant difference between the performance obtained with the dichotic presentation (low-high) and monaural presentation for 8 and 6 channels. But similar difference was not observed for the 12 channel presentation. The results also showed that presenting odd or even channels alone to normal hearing subjects was not sufficient. Post-hoc Fishers LSD tests showed that there was a significant difference ($p<0.05$) between the one-ear performance and the dichotic performance on sentence recognition, suggesting that normal hearing subjects were able to integrate the information from the two ears.

- Examination of the results obtained for the vowel recognition showed a significant effect of spectral resolution on the vowel recognition. Comparing the performance of the dichotic presentation with the performance of left ear only and right ear only clearly showed the ability of the subjects to integrate the information provided to the two ears.

- Examination of the similar test results for the consonants showed that spectral resolution had no significant effect [$F(2,16)=4.79$, $p=0.34$] on the consonant recognition. Consonants were fused accurately with both dichotic presentations regardless of the spectral resolution.

- The performance obtained with the interleaved dichotic condition was not significantly different from the performance obtained monaurally when the speech materials were processed through 12 channels. For bilateral patients receiving a large number of channels (>12), dichotic stimulation might therefore provide an advantage over unilateral stimulation, in that it could potentially reduce possible channel interactions since the electrodes could be stimulated in an interleaved fashion across the ears without sacrificing performance. For patients receiving only a small number of channels of information, dichotic (electric) stimulation might not produce the same level of performance as unilateral (monaural) stimulation, at least for sentences presented in the low-high dichotic condition.

- Results of sentence recognition with cochlear subjects in quiet (p = 0.28) clearly showed that the subjects were able to integrate the information provided to both the ears.

- The experiments in noise showed that the noise level in the speech degrades the fusion capability of the subjects. The comparison of the results obtained from dichotic and diotic condition clearly showed a significant drop in recognition of the sentences presented dichotically.

Preliminary experiments were performed to investigate the effects of beamforming on speech recognition in noise. The speech samples were processed through Griffiths-Jim beamforming algorithm and then compared with the unprocessed noisy speech. Results indicated no significant benefit with beamforming. More subjects are needed however to draw any reliable conclusions with beamforming.

REFERENCES

[1] R. Tyler, J. Preece, and M. Lowder. The iowa audiovisual speech perception laser videodisc. Laser Videodisc and Laboratory Report, Dept. of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinincs, 1987.

[2] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25, 1953.

[3] D. E. Broadbent and P. Ladefoged. On the fusion of sounds reaching different sense organs. *J. Acoust. Soc. Am.*, 29, 1957.

[4] T. C. Rand. Dichotic release from masking for speech. *J. Acoust. Soc. Am.*, 55, 1974.

[5] J. E. Cutting. Auditory and linguistic process in speech perception: Inferences from six fusions in dichotic listening. *Psychol. Rev.*, 83, 1976.

[6] C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Q. J. Exp. Psychol.*, 33A, 1981.

[7] T. Lunner, S. Arlinger, and J. Hellgren. 8-channel digital filter bank for hearing aid use: Preliminary results in monaural, diotic and dichotic modes. *Scand. Audiol. Suppl.*, 38, 1993.

[8] P. C. Pandey, D. S. Jangamashetti, and A. N. Cheeran. Binaural dichotic presentation to reduce the effect of temporal and spectral masking in sensorineural

hearing impairment. In *142nd Meeting of the Acoustical Society of America*, 2001.

[9] P. Loizou, M. Dorman, and V. Powell. The recognition of vowels produced by men, women, boys and girls by cochlear implant patients using a six-channel cis processor. *J. Acoust. Soc. Am.*, 103, 1998.

[10] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler. Acoustic characteristics of american english vowels. *J. Acoust. Soc. Am.*, 97, 1995.

[11] M. Nilsson, S. Soli, and J. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95, 1994.

[12] P. Loizou, M. Dorman, and Z. Tu. On the number of channels needed to understand speech. *J. Acoust. Soc. Am.*, 106, 1999.

[13] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Am.*, 27, 1955.

[14] P. Loizou, O. Poroy, and M. Dorman. The effect of parametric variations of cochlear implant processors on speech understanding. *J. Acoust. Soc. Am.*, 108, 2000.

[15] D. Lawson, B. Wilson, M. Zerbi, and C. Finely. Fourth quarterly progress report. Center for Auditory Prosthesis Research Research Triangle Institute, 1999.

[16] R. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270, 1995.

[17] M. Dorman, P. Loizou, and D. Rainey. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J. Acoust. Soc. Am.*, 102, 1997.

[18] R. P. Lippmann. Accurate consonant perception without mid-frequency speech energy. *IEEE. Trans. Speech Audio Process*, 4(1), 1996.

[19] K. Kasturi, P. Loizou, and M. Dorman. The intelligibility of speech with holes in the spectrum. *J. Acoust. Soc. Am.*, 112, 2002.

[20] M. Dorman, P. Loizou, A. J. Spahr, E. S. Maloff, and S. V. Wie. Speech understanding with dichotic presentation of channels: Results from acoustic models of bilateral cochlear implants. In *Conference on Implantable Auditory Prosthesis, Asilomar*, 2001.

[21] P. E. Lyregaard. Frequency selectivity and speech intelligibility in noise. *Scand. Audiol. Suppl.*, 15, 1982.

[22] B. Franklin. The effect of combining low and high-frequency bands on consonant recognition in the hearing impaired. *J. Speech Hear. Res.*, 18, 1975.

[23] S. Turek, M. Dorman, J. Franks, and Q. Summerfield. Identification of synthetic /bdg/ by hearing impaired listeners under monotic and dichotic formant presentation. *J. Acoust. Soc. Am.*, 67, 1980.

[24] P. Loizou, O. Poroy, M. Dorman, and T. Spahr. Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution. *J. Acoust. Soc. Am.*, 108, 2000.

[25] P. Loizou and O. Poroy. Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners. *J. Acoust. Soc. Am.*, 110, 2001.

[26] M. Dorman and P. Loizou. The identification of consonants and vowels by cochlear implants patients using a 6-channel cis processor and by normal hearing listeners using simulations of processors with two to nine channels. *Ear Hear*, 19, 1998.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. The darpa timit acoustic-phonetic continuous speech corpus cdrom. Technical report, National Institute of Standards and Technology, 1993.

[28] R. J. M. VanHoesel and R. S. Tyler. Speech perception, localization, and lateralization with bilateral cochlear implants. *J. Acoust. Soc. Am.*, 113(3), 2003.

[29] J. E. Greenberg and P. M. Zurek. Evaluation of an adaptive beamforming method for hearing aids. *J. Acoust. Soc. Am.*, 91(3), 1992.

[30] W. G. Gardner and K. D. Martin. Hrtf measurement of a kemar. *J. Acoust. Soc. Am.*, 97(6), 1995.

[31] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, January 1982.

VITA

ArunVijay Mani was born on April 20, 1980, the son of R. Mani and M. Vathsala. He finished his high school at St.John's Vestry Higher Secondary School, Trichy, India in 1997. In the same year he was admitted to the Bachelor program in the Electronics and Instrumentation Department in Government College of Technology (under the Barathiyar University), India. He was awarded the Bachelor of Engineering in Electronics and Communication in April 2001. In August 2001, he was admitted to the Masters program in Electrical Engineering (specialization in Telecommunication), at the University of Texas at Dallas. From Spring 2002 he has been working in Speech Processing and Cochlear Implant Lab. His research mainly concerns the speech recognition by the Cochlear Impalnt listeners. He was a Teaching Assistant in the Electrical Engineering Department during Spring 2002. He is currently in the final graduating semester of his MSEE program.