EXPERIMENTAL EVALUATION OF IMPROVEMENT IN SPEECH PERCEPTION WITH CONSONANTAL INTENSITY AND DURATION MODIFICATION

Thesis

Submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy

by

T. G. Thomas

DEPARTMENT OF ELECTRICAL ENGINEERING
Indian Institute of Technology
Powai, Bombay - 400 076

APRIL 1996

APPROVAL SHEET

Thesis entitled: EXPERIMENTAL EVALUATION OF IMPROVEMENT IN SPEECH PERCEPTION WITH CONSONANTAL INTENSITY AND DURATION MODIFICATION

by T. G. Thomas

is approved for the degree of DOCTOR OF PHILOSOPHY

Examiners

S S Bhave
(S.S. Bhave)

(P.G. Poonacha)

Supervisors

PCPandey
(P.C. Pandey)

S. D Agashe
(S.D. Agashe)

Chairman

S L Dhingra
05/04/96
(S.L. Dhingra)

Date: 03/04/96

Dedicated to the memory of

my beloved father

who had been a great

source of inspiration

for higher learning

CONTENTS

Page

APPENDICES

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my thesis supervisors, Dr. P. C. Pandey and Dr. S. D. Agashe, for their constant encouragement, invaluable guidance, and moral support throughout the course of this research work. They have helped me in more ways than one, both in academic and non-academic matters, and I am extremely beholden to them for the same.

I thank Dr. P. G. Poonacha, who as my internal examiner, showed interest and spared no pains in evaluating my work and giving useful comments. I am also grateful to Dr. Girish Kumar and Dr. T. Anjaneyulu for their comments.

Any working place is full of excitement, tensions, and depressions. It is here that the presence of friends helps in providing the necessary support and relief. I thank my lab-mates Joshi, Nilesh, Anand, Prashant, Vipul, Harshal, Parmar, Sapre, Khambete, Prasad, and Ashok for their help. In fact, this thesis would not have seen the light of day if some of them had not literally "lent their ears" as subjects.

I thank Shri. Apte and Shri. Misquitta for their help as also the non-teaching staff of the Instrumentation Lab, namely, Shri. Wagh, Shri. Iyer, and Shri. Mantode for their help in various ways.

I thank the office staff of the Department for their cooperation.

I am extremely grateful to the authorities of the University College of Engineering in particular and Osmania University in general for sponsoring me for Ph. D. under the Quality Improvement Programme (Q. I. P.). The help of Prof. R. Narasimhan of the Electrical Engineering Department is gratefully acknowledged. I also thank my colleagues in the Electronics and Communication Engineering Department for their encouragement.

I am grateful to Mrs. Rekha Roy, Director, Ali Yavar Jung National Institute for the Hearing Impaired, Bombay for readily providing the facilities at her Institute and also for the use of the library. I also thank Dr. Nagaraja and Dr Ravi Shankar for evincing interest in my work and Shri. Venugopal for conducting the hearing screening test and headphone calibration work.

I thank the Q. I. P. Coordinator, Dr. T. S. Raghunathan, and his office staff for their help in many ways. I also thank all my neighbours at the Q. I. P. Building who made the place a home away from home.

No words will express my feelings of affection and respect towards my wife Molly and son Tushar for going through periods of turbulence with fortitude. I have denied them many happy moments but they strove to give me joy, support, and encouragement at the home front.

I thank the Lord God Almighty for giving me the necessary strength and determination to complete my work.

In the true IIT-ian spirit, I can sum up this phase of my life as follows:

"It was the best of times, it was the worst of times."

Thomas

T. G. Thomas

NOMENCLATURE

AF	amplitude of frication
AH	amplitude of aspiration
AL	alveolar
AV	amplitude of voicing
BD	burst duration
B_i	i^{th} formant bandwidth
CD	consonant duration
CV	consonant-vowel
CVR	consonant-to-vowel intensity ratio
CV6	consonants /p, t, k, b, d, g/ in the consonant-vowel context of the vowel /a/
CV9	unvoiced consonants /p, t, k/ in the consonant-vowel context of the vowels /a, i, u/
C/V	consonant-to-vowel
D/A	digital to analog
F_0	fundamental frequency of the voice pitch
F_i	i^{th} formant frequency
FTD	formant transition duration
LA	labial
LDL	loudness discomfort level
MLP	mean logarithmic probability
R_s	response (or articulation) score
SD	standard deviation
SNR	signal-to-noise ratio
SPL	sound pressure level
T_0	voice pitch period
VC	vowel-consonant
VC6	consonants /p, t, k, b, d, g/ in the vowel-consonant context of the vowel /a/
VC9	unvoiced consonants /p, t, k/ in the vowel-consonant context of the vowels /a, i, u/
VE	velar
VOT	voice onset time

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Speech is a complex signal with temporal and spectral variations. However, only certain aspects of the acoustic signal seem to be relevant to the listener for perceiving the phonetic dimensions of speech. These relevant attributes for speech are called acoustic correlates or acoustic cues. In perceptual studies, these acoustic cues are manipulated systematically to determine their importance in terms of phoneme discrimination and identification [1, 2, 3]. Enhancement of the more important cues is expected to facilitate improved phoneme recognition by the hearing impaired.

According to a survey conducted by the National Sample Survey Organisation in 1981, there were an estimated 3.36 million people with hearing disabilities in India out of a total population of about 680 million [4]. In this context, there were about 19 million people with hearing impairments in the USA in 1985 out of a total population of about 240 million [5].

Acoustic amplification is commonly used in order to enhance speech recognition by hearing-impaired people. However, for people with sensorineural hearing impairments, the solution is a bit more complex. There is convincing evidence that hearing impairment of cochlear (i.e., sensory) origin results in aberrations in intensity perception, and poorer-than-normal temporal and frequency resolutions [6, 7]. The frequency-dependent shifts in hearing thresholds, without any significant shift in the threshold of discomfort, reduces the dynamic range considerably. This causes an abnormally rapid growth of loudness when the stimulus level is increased. Protection against excessive amplification has been provided by means of peak clipping or amplitude compression; however, both of these have failed to yield significant improvements in speech intelligibility.

Another possibility is a form of acoustic recoding known as frequency transposition or frequency lowering. The residual hearing in advanced cases of sensorineural impairment is often restricted to frequencies up to about 1 kHz. Frequency lowering shifts the inaudible speech energy in the high frequencies into the low-frequency region, where it produces distinctly audible cues. Though frequency lowering works well with voiceless fricative sounds [8], its success in recoding sounds that contain information in both the low and high frequencies have not been proved. The improvements that have been reported have been limited to specific applications such as speech training and considerable amount of training was required by the subjects to learn the new code.

A number of research studies have aimed at identifying and enhancing important speech characteristics that are distorted, weak, or imperceptible for individual hearing-aid users. One promising scheme for enhancing the speech signal to improve intelligibility is based on studies of speaking clearly for the hearing impaired [9]. Studies of the difference between "clear" and "conversational" speech suggest that it may prove beneficial to attend to the temporal characteristics of speech Acoustic differences have been identified in "clear" speech which can yield consistent gains in intelligibility for the hearing impaired. Two specific parameters are the consonant-to-vowel (C/V) intensity ratio and consonant duration which are found to increase in the case of "clear" speech.

Subsequent studies have evaluated the intelligibility of natural speech artificially transformed to "clear" speech by altering either one or both of these acoustic parameters [10, 11]. While all the studies reported improvements in recognition for C/V intensity ratio modification, albeit by different amounts, the effect of consonant lengthening has been indeterminate. This may have been because of differences in stimuli, subject background, effects of signal processing, etc. Furthermore, the acoustic segments associated with consonant phonemes are found to increase in a nonuniform manner and such changes cannot be aptly simulated by artificially transforming natural speech to "clear" speech.

The above studies on the effect of C/V intensity ratio modification and lengthening consonant duration used naturally uttered speech syllables as stimuli. By using synthetic speech, it would be possible to manipulate the spectral, temporal, and intensity characteristics independently. The segmentation of individual phonemes would be easier and it would be possible to alter the various acoustic segments independently. Hence, it was decided to use a speech synthesizer to generate the stimuli used in the present investigation.

1.2 RESEARCH OBJECTIVES

The importance of "clear" speech as a means of improving speech recognition by the hearing impaired was seen in the last section. The advantages of using synthesized speech material for studies in speech perception were also indicated.

The aim of the present investigation was to study the extent of consonant-to-vowel intensity ratio (CVR) modification and consonant duration (CD) modification that would be helpful in speech perception by the hearing impaired. Synthesized nonsense syllables involving stop consonants were used as the test stimuli. Listening tests were conducted on normal-hearing persons both in the quiet and under simulated hearing impairment conditions. Hearing impairment was simulated by mixing the stimuli with broadband masking noise. Experiments were performed to study the following aspects:

1) The effect of varying CVR on the recognition of the consonants /p, t, k/ in the /consonant-vowel/ (CV) and /vowel-consonant/ (VC) contexts of the vowels /a, i, u/. Information transmission analysis [12] was used to study the effect of CVR modification on the overall information transmission and on the transmission of consonant and vowel features. It was also to be investigated whether vowel recognition was impaired with CVR modification.

2) The effect of varying CVR on consonant recognition when the stimuli included both voiced and unvoiced stops, namely, /p, t, k, b, d, g/, in the CV and VC contexts of the vowel /a/. The effect of CVR modification on the transmission of

both place and voicing features under different SNRs was to be investigated.

3) The effect of CD modification on recognition of the consonants /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/. Stimuli were generated in which the formant transition duration, voice onset time, and burst duration were independently altered to study their individual effects on consonant recognition. The effect of altering these acoustic segments on the transmission of place and voicing features was to be investigated using information transmission analysis.

1.3 THESIS OUTLINE

Chapter 2 describes the characteristics of an impaired auditory system and their effect on speech perception. This is followed by an overview of electroacoustic hearing aids and various signal processing strategies proposed for improving speech perception. The characteristics of "clear" speech and "conversational" speech are described in Chapter 3. Some of the reported methods of improving perception of natural speech by artificial transformation to "clear" speech are then considered and the need for the present investigation is justified.

Chapter 4 describes the properties of stop consonants in order to better understand the nature of the stimuli used in the present investigation. This is followed by a brief discussion on speech synthesis and the method adopted in synthesizing the stop consonant stimuli. Chapter 5 presents the experimental methodology adopted. Some of the techniques for analyzing and summarizing confusion matrix data are reviewed and the choice of the techniques used for analysis of test results is also discussed.

Chapter 6 presents the results and implications of recognition experiments performed on different sets of synthetic speech stimuli which have been modified by increasing the C/V intensity ratio. The results and implications for experiments performed on stimuli with increased consonant duration are presented in Chapter 7. Chapter 8 gives a summary of the conclusions drawn from the present investigations and suggestions for future work.

The appendices provide supplementary information and data. The spectrographic analysis of speech waveforms is described in Appendix A. Appendix B gives the parameter tracks of some of the stimuli used in the experiments. Details of the hardware and programs for signal handling and experimental control in the computerized listening test administration are presented in Appendix C. Appendix D gives a brief description of the programs for the analysis of confusion matrices obtained from the speech tests. Appendix E considers the effect of sample size on the results obtained. Appendix F describes the procedure adopted for the electroacoustic calibration of the DR-59 headphone which was used in the experiments. The subject data and test instructions given to the subjects are included in Appendix G and Appendix H respectively.

CHAPTER 2
SPEECH PERCEPTION BY THE HEARING IMPAIRED

2.1 INTRODUCTION

This chapter first describes the characteristics of the impaired auditory system and their effect on speech perception. This is followed by an overview of the electroacoustic hearing aids and signal-processing strategies that have been investigated by different researchers for improving speech reception using these aids.

2.2 THE IMPAIRED AUDITORY SYSTEM

Hearing impairments may be classified into four major categories [13, 14]: conductive loss, sensorineural loss, central loss, and functional deafness. Conductive losses result from dysfunction of the ear canal or middle-ear structures, so that less acoustic energy reaches the auditory receptors in the cochlea. Sensorineural losses are due to reduced sensitivity of the neural receptor mechanisms themselves. Most conductive losses are successfully treated through medical intervention. Sensorineural losses are generally not amenable to medical intervention, and patients need to use aids for speech reception. Central impairment may be due to damage to the auditory cortex by cerebral hemorrhage, meningitis, etc. It is not necessarily accompanied by a decrease in auditory sensitivity but tends to manifest itself in varying degrees through a decrease in auditory comprehension. Functional deafness has no known organic involvement and the causes are psychological rather than physiological. All future discussion here pertains to sensorineural hearing impairment, unless otherwise stated.

Sensorineural losses can alter peripheral processing in each of the physical dimensions of the speech signal — intensity, time, and frequency [6, 7]. The aberrations in intensity perception may be in the form of frequency-dependent shifts in hearing thresholds, a reduced dynamic range, and an abnormally rapid growth of loudness with increasing stimulus level. The time resolution for detecting acoustic events gets degraded, and it is often accompanied by forward and backward masking of weaker speech sounds by more intense sounds [15]. The poor frequency resolution results in degradation of perception of speech sounds on the basis of spectral characteristics. While discussing each of these characteristics, it would be interesting to note how the perception of subphonemic features of speech get affected by them.

Fig. 2.1 shows the hearing thresholds for typical sensorineural impairments (dashed curves A, B, and C). The hatched portion indicates the area within which typical speech spectra lie. It can be seen that while the lower-frequency components of speech are detectable, the higher-frequency components are not. English consonants are uniquely classified by three features: voicing (whether or not the laryngeal voicing source is active during consonant production), manner of articulation (the type of vocal tract gesture made) and place of articulation (the point in the vocal tract of maximum constriction during production). In a study of feature recognition by normal-hearing listeners, Miller & Nicely [12] showed that as more and more high-frequency information is filtered out of the speech signal, the number of place errors increases markedly, while perception of voicing was largely unaffected. On the basis of this and other studies, the significance of the shape of the speech spectrum can be summarized: The high-power, relatively low-frequency portion of the spectrum around 500 Hz contains first formant vowel energy, as well as information about consonant voicing and manner. The mid-frequency region between 500 and 2000 Hz contains second formant energy, which is important for both the identification of vowels and of the place of articulation of consonants. The high-frequency, low-power end of the spectrum represents upper formant energy and consonant noise associated with stop and fricative consonants.

As seen from Fig. 2.1, the impaired auditory system is accompanied by a reduced dynamic range. The degree of impairment increases from mild-to-moderate for curve A to severe for curve B, and profound for curve C. Curve D shows the normal threshold of hearing. The dashed line at the top is the loudness discomfort level (LDL) which, being fairly similar for both normal hearing and sensorineurally hearing-impaired persons, is represented by a single curve. The distance between the threshold curve and the LDL curve is defined as the effective dynamic range of the auditory system. The area enclosed by these two curves is known as the residual hearing area. The residual hearing area becomes progressively smaller with increasing hearing loss. The dynamic range at high frequencies may be only a few dBs for those with severe losses. A consequence of drastically reduced dynamic range in some cases is loudness recruitment which involves abnormally rapid growth of loudness with stimulus level.

The temporal variations in level in normal conversational speech are estimated to cover a range of at least 30 dB [16] which is large compared with the available dynamic range for most hearing-impaired listeners. A sound can be heard only if its level exceeds the threshold of audibility. Acoustic amplification helps in raising many of the weaker sounds of speech into the residual hearing area. Unfortunately, there is a limit to the amount of amplification that can be provided. Because of the reduced dynamic range of the impaired ear, the amplification needed to make the weak sounds of speech audible will also make the intense sounds uncomfortably loud. Continued exposure to sounds above the loudness discomfort level can damage whatever residual hearing exists.

Temporal resolution refers to the minimum time required to resolve acoustic events. It is most popularly measured by means of a gap-detection task, in which listeners have to detect the presence of a temporal gap in a burst of noise. Hearing-impaired listeners often show poorer-than-normal temporal resolution [17]. However, this does not prevent listeners with up to moderately severe hearing impairment from identifying stimuli that differ in voice onset time (VOT) [18]. Listeners with severe hearing impairment do exhibit a deficit in resolution of VOT, but that is when they are forced to process the cues for place of articulation as well as for voicing.

Hearing impaired listeners have been shown to be at least as sensitive as normal listeners to relatively small changes in vowel duration when vowel duration is one of the cues to final consonant voicing [19]. Hannley & Dorman [20] have reported that hearing impaired listeners have more difficulty in identifying place of articulation for stop consonants when the cue is a falling second formant transition than when the cue is a rising second formant transition.

Frequency resolution of the ear refers to its ability to resolve the frequency components of a complex signal which are temporally coincident. Poor frequency selectivity is manifest as greater-than-normal upward and downward spread of masking [20]. It is caused by the abnormally wide bandwidths of the internal auditory filters in hearing-impaired listeners [14]. The effect of poor frequency selectivity is to reduce the difference in the amplitude of the spectral peaks and troughs in the internal auditory representation of the acoustic signal. The resulting loss of sharply defined spectral peaks may lead to uncertainty in the location of spectral maxima. It has been shown by Leek *et al.* [21] that while normal-hearing listeners can identify vowels with spectral peaks only 2 dB above the level of the troughs, listeners with moderate hearing impairment need at least 7 dB. However, the identification of vowels by hearing-impaired listeners is still good. This is because vowels, in natural speech, are characterized by peak-to-trough differences that are at least 8 to 10 dB.

The importance of consonants, relative to vowels, as the major carriers of information has been recognized even by early investigators [22]. In contrast to vowels, stop consonants are not well identified by hearing-impaired listeners. The acoustic information that identifies vowels and stops is quite different. The stops are characterized not only by the location of spectral peaks, but also by the change in frequency of those peaks over the first 20 to 50 ms of the signal, by the tilt of the amplitude-frequency spectrum at signal onset, and by how long the spectrum remains constant following signal onset. For normal-hearing listeners, spectral cues determine identification performance. For hearing-impaired listeners, spectral cues have less influence while tilt and time cues have more [7]. Hence, altering the temporal properties of speech signals for better intelligibility appears to be a promising prospect.

On the basis of several studies reported in the literature, certain broad conclusions regarding consonant recognition by hearing-impaired listeners may be made:

1) Vowel identity is generally well perceived in both the /consonant-vowel/ and /vowel-consonant/ contexts [23, 24].
2) Consonant recognition is better in syllable-initial position than in syllable-final position [9, 25]. However, certain studies have reported better recognition of consonants in syllable-final position over those in syllable-initial position [26, 27].
3) Confusions between consonants differing in manner are less common than place errors [3].
4) Voiced consonants are identified better than unvoiced consonants [28, 29]. However, Gordon-Salant [11] reported

higher recognition of unvoiced consonants over voiced consonants.

2.3 ELECTROACOUSTIC HEARING AIDS

The benefits of amplifying sound in order to help alleviate the effects of hearing impairment have been known since early times. The early "amplifiers" included cupping one's hand behind the ear and the ear trumpet. The modern history of the hearing aid began with the invention of electrical amplification. The first commercially produced electrical hearing aid appears to have been the *Akouphone*, a carbon-microphone hearing aid invented by Hutchinson in 1902 [13]. Amplification was achieved in the transduction process from acoustic to electrical signals in this hearing aid.

The development of hearing aids has greatly benefitted from advances in communications technology over the years. Hearing aids were among the first products to make use of the miniaturized vacuum tube, the transistor, integrated circuits, and miniature electret microphones. The conventional electroacoustic hearing aids typically consist of a microphone, electronic filter, controls for adjusting the amplification and overall shape of the frequency response, circuits for limiting the amplified signals to a comfortable or safe level, an earphone (receiver), and a battery that serves as the power source. In addition, there are various acoustic components, such as flexible tubing and an earmould, for coupling the output of the receiver to the external ear canal [30]. The electroacoustic hearing aid has been improved in a variety of ways in the past three decades. During this period, two other classes of speech-perception aids have also been developed: (1) surgically implanted cochlear prostheses [31] and (2) sensory-substitution systems, including vibrotactile and electrocutaneous stimulating devices [32].

The CHABA study [5] on speech-perception aids for the hearing-impaired observed that individuals whose speech-frequency thresholds show mild-to-severe hearing loss (25-90 dB) will receive more usable speech-waveform detail through the electroacoustic hearing aid than from the other two classes of aid that transform sound into nonacoustic stimulation. For those with speech-frequency losses in excess of 115 dB, no significant help will be derived from acoustic amplification. Such persons are clearly candidates for either a cochlear implant or for one of the sensory-substitution aids. For those with losses in the region 90-115 dB, a region in which the listener may receive small but significant benefits from acoustic amplification, the choice of aid is highly dependent on the needs and expectations of the impaired listener and on the person making the recommendation.

The largest wearable electroacoustic hearing aid is the body-worn aid, in which the electronic components are housed in a body-worn case, the amplified signals being delivered by wire to a receiver mounted in the ear. These are used by persons requiring very high-powered acoustic output and when cosmetic factors are of secondary importance.

In the behind-the-ear (BTE) hearing aid, the electronic components are housed in a small elliptical case that fits behind the ear. The acoustic signals from the receiver are delivered to the ear canal by means of a flexible acoustic tube terminating in an earmould. These aids provide a more natural signal through the use of ear-level microphones, and they also have a better cosmetic appeal.

The in-the-ear (ITE) hearing aid goes one step further in size reduction. All the components are housed in a small plastic case that fits into the outer portion of the external ear canal.

The smallest hearing aid is the in-the-canal (ITC) aid, which fits entirely in the ear canal and is the least conspicuous of the lot. The BTE has been replaced by the ITE or ITC aids as the instrument of choice in western countries. This appears to be because of the small size (and hence low visibility) of the ITE and ITC aids. However, they are limited in terms of their acoustic power output and are appropriate only for persons with mild or moderate hearing losses.

Other commercially available acoustic amplification systems for hearing-impaired people include special-purpose amplifiers for use on the telephone, radio, and television as well as personal FM transmission systems for use in noisy settings like classrooms or auditoriums [33].

2.4 SPEECH PROCESSING FOR HEARING AIDS

Sensorineural hearing impairments are mainly characterized by increase in hearing thresholds and reduction in dynamic range. Whereas linear amplification is addressed to the problem of elevated thresholds, amplitude compression, a form of nonlinear processing, is concerned with reducing the range of speech level variation to match the reduced dynamic range of the impaired ear. There are three distinct types of amplitude compression functions. Limiting is a very rapid form of compression intended to protect the listener from sound intense enough to cause discomfort or pain. Automatic volume control (AVC) is a relatively slow acting form of compression which serves to control relatively long-term variations in speech levels. Multiband compression systems have independent AVC circuits for each frequency band [34].

Attempts at improving intelligibility for impaired listeners by employing multiband compression have met with limited success. Although [35] reported substantial advantage for a two-channel compression system over linear amplification, subsequent studies [36, 37] failed to confirm this result. In particular, Lippmann *et al.* [36] found multiband compression no better than linear amplification with high-frequency emphasis appropriate to the impairment. A more recent study [38] with a multiband compression system seems to be holding some promising results.

Experimental evaluations have shown that, as a protective device, compression limiting is superior to simple peak clipping. There is less distortion of the amplified speech and, therefore, speech intelligibility is reduced less by compression limiting than by peak clipping [39]. Although amplitude compression and peak clipping provide protection against excessive amplification, both of these have failed to yield significant improvements in speech intelligibility [5].

In the case of severe or profound sensorineural hearing impairment, it is quite common to find some hearing in the low frequencies (up to about 1 or 2 kHz). A form of acoustic recoding known as frequency transposition or lowering shifts the inaudible speech energy in the high frequencies into the low-frequency region, where it produces distinctly audible cues. Over the years, a number of frequency-lowering systems have been developed using both selective and total-waveform lowering (see [40], for a review). Experimental evaluations of frequency lowering for hearing-impaired people have yielded mixed results. Johannson [8] evaluated his transposer system and obtained improved discrimination of fricatives and other phonemes by profoundly hearing-impaired children. Ling [41] reported a series of experiments using the Johannson-type frequency transposer in which no significant advantages over conventional amplification were obtained for either speech reception or speech training. However, Foust & Gengel [42] reported significant advantages over conventional amplification in the speech discrimination ability of individual subjects, but only after a fair amount of training. Reed *et al.* [43] have studied pitch-invariant frequency lowering, using nonuniform frequency compression of the short-term spectral envelope, but without any significant improvement in speech discrimination performance.

With the increasing availability of sophisticated digital signal-processing techniques, attention has been directed towards techniques for enhancing specific phonetic features. For the case of vowels in noise, improving the salience of the formants by decreasing the acoustic energy in the valleys between spectral peaks resulted in improved intelligibility [21].

Most previous research directed towards improved speech intelligibility for the hearing impaired has focussed on the signal processing of conversational speech. Studies [9, 44] have shown that it is possible to create more intelligible speech for the hearing impaired by instructing speakers to talk more clearly. Picheny tested impaired listeners on words in nonsense sentences spoken both "clearly" and "conversationally" and analyzed the acoustic differences between the two types of materials. Test results showed a clear and robust intelligibility advantage of 17% for clear speech.

Studies of the differences between "clear" and "conversational" speech suggest that it may be particularly fruitful to attend to the temporal characteristics of speech. The next chapter discusses the nature of these differences, the results of reported studies artificially transforming natural speech to "clear" speech, and the motivation behind the present investigation.

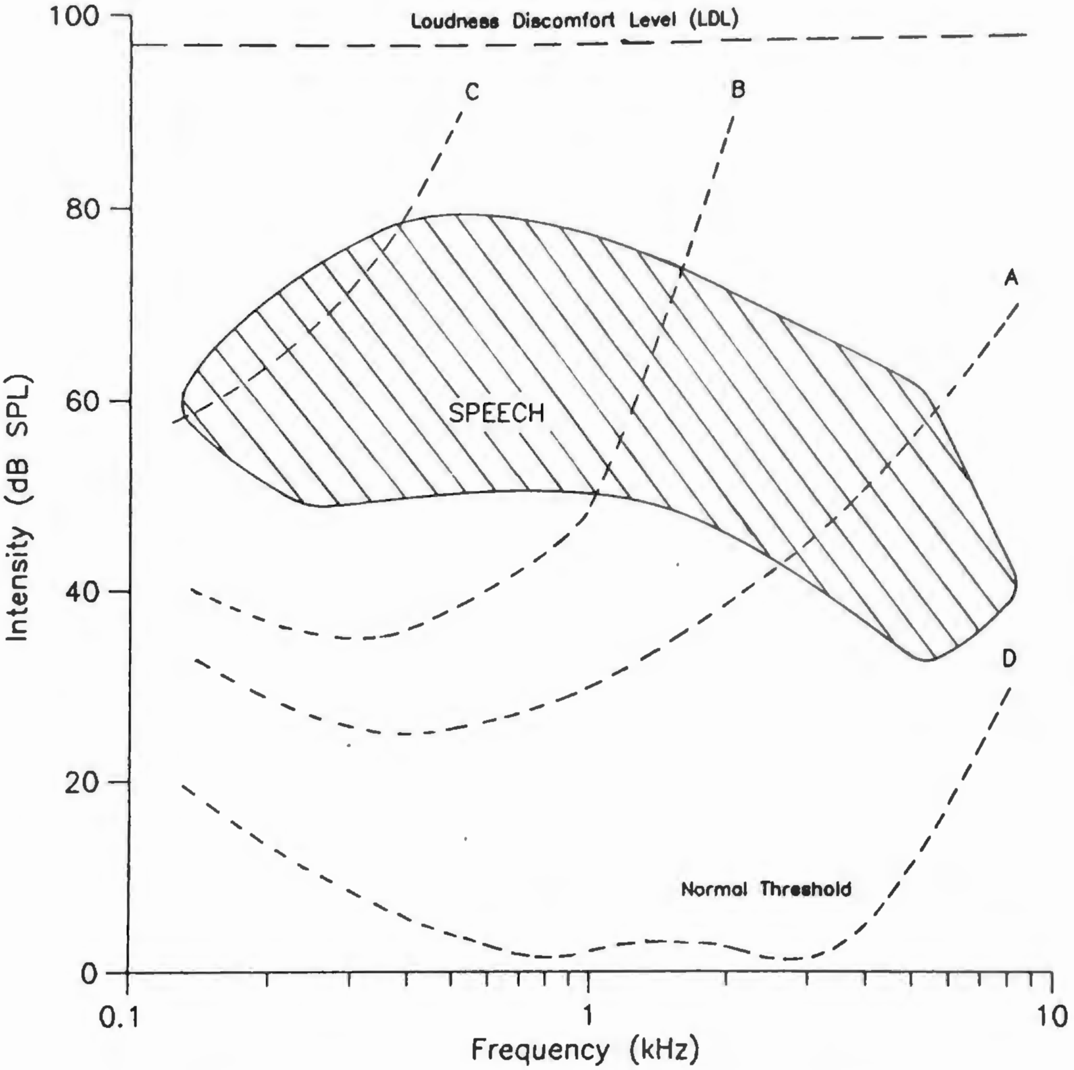


FIG. 2.1 Illustrating the dynamic range of the impaired auditory system. Curves A, B, C are the hearing thresholds for typical sensorineural impairments. Hatched portion indicates area within which typical speech spectra lie. (Adapted from [5] and [62])

CHAPTER 3
SPEECH ENHANCEMENT BASED ON THE PROPERTIES OF CLEAR SPEECH

3.1 INTRODUCTION

This chapter first describes the characteristics of "conversational" speech and "clear" speech. Some of the reported methods of improving intelligibility of natural speech by artificial transformation to clear speech are then studied. Finally, the motivation behind the present investigation is discussed.

3.2 CONVERSATIONAL SPEECH AND CLEAR SPEECH

"Conversational" speech is the speech which occurs between people in normal, everyday situations. In contrast, "clear" speech can be defined as that speech which occurs when one is trying to improve communication in a difficult situation, as when speaking in a noisy environment or to a hearing impaired person [9]. The increased clarity may be obtained by changing the conversational context, sentence structure, vocabulary, speaking rate and stress, pronunciations of individual words and speech sounds, and vocal effort.

Picheny [9] established that clear speech was more intelligible than conversational speech to hearing-impaired listeners, and has noted that the identification of the acoustic characteristics associated with intelligible speech could be of considerable value for hearing aid research. Once they are identified, it may be possible to enhance those characteristics through signal processing to improve speech recognition by hearing-impaired individuals. He conducted tests with words in nonsense sentences spoken both "clearly" and "conversationally" and analyzed the acoustic differences between these two types of materials. Materials were presented to five impaired listeners at three intensity levels using two frequency-gain characteristics. Test results showed that the intelligibility of clear speech was 17% higher. Substantial improvements were obtained for all listeners, all (three) speakers, and all levels. In addition, the improvements appeared to be roughly independent of word position in the sentence and phoneme class. Acoustic analysis indicated that there were large differences in the temporal characteristics of the two types of speech materials. In clear speech, the speaking rate was reduced by a factor of approximately two; significant increases were observed in the number and duration of pauses, the duration of most speech sounds, and the range of fundamental frequency variation. Analysis of phonetic characteristics showed no substantial change in long-term spectra, a non-uniform increase in the consonant-vowel intensity ratio and a variety of changes in individual consonants and vowels.

The following sections consider some of the reported studies on speech enhancement based on manipulating the consonant/vowel intensity ratio and the phoneme durations.

3.3 CONSONANT-TO-VOWEL INTENSITY RATIO MODIFICATION

The consonant-to-vowel (C/V) intensity ratio has been defined and measured differently by different investigators, but in general it refers to the difference in decibels between either the power or the energy of the consonant and that of the adjoining (preceding or following) vowel [45]. Several earlier studies that have considered intelligibility differences among talkers seem to support the view that C/V ratio is important. House *et al.* [44] measured the C/V ratios for two talkers differing in intelligibility on the Modified Rhyme Test. They found that the more intelligible talker had C/V ratios 2-4 dB higher than the less intelligible talker. One of the factors studied by Picheny *et al.* [46, 47] was the intelligibility of key words within conversationally spoken sentences for three different talkers. The speech of the talker who was most easily understood by the five hearing-impaired listeners also had the highest average consonant intensity. Since the stimuli were calibrated with respect to syllable peaks this implied that the speech of this talker had the highest C/V ratios.

Several researchers have conducted investigations using test materials whose C/V ratios were manipulated using digital signal processing. Ono *et al.* [10] found that nonsense syllable recognition was improved with increased C/V ratio. Performance improvements of about 10 to 15% have been reported for consonants amplified by 10 to 21.5 dB above their natural level relative to vowels in words [48] or nonsense syllables [49]. The listeners employed in the above studies had generally mild and moderate impairments. A similar improvement was reported [11] for young and elderly normal-hearing subjects in the presence of background

[illegible]

In comparison to listeners with moderate or mild hearing impairments, severely hearing-impaired subjects have not obtained comparable increases in performance from consonant amplification ([illegible] increased C/V ratio). Revoile *et al.* [50] reported identification at chance level for amplified stop consonants for listeners with severe to profound hearing impairments. However, the same listeners showed considerably improved recognition for distinctions of consonant voicing (91% for amplified stops versus 68% for unmodified stops). A more recent study [51] considered the effect of C/V ratio modification on amplitude envelope cues by employing normal listeners with simulated profound hearing loss. Profound hearing loss was simulated by using stimuli consisting of pink noise modulated by the amplitude envelopes extracted from vowel-consonant-vowel utterances. In the process spectral information was minimized as it is expected to be in cases of profound hearing loss. The consonant portion of each utterance was amplified by 10 dB. Recognition performance in the amplified-consonant condition was reduced for some consonants like glides and voiced fricatives. However, for other consonants like stops and unvoiced fricatives, consonant amplification increased recognition. No significant change in recognition was observed for amplified affricates and nasals.

It should be noted that amplitude compression (discussed in the previous chapter), which tends to increase C/V ratio as one of its effects, has yielded at best mixed results with hearing-impaired listeners. Even carefully controlled multiband compression has yielded ambiguous findings. It has been suggested [48] that the degradation of intelligibility due to multiband compression was possibly caused by distortion of amplitude relationships within a phonetic segment. Such a situation does not exist in the studies cited above that specifically manipulated consonant intensity.

3.4 DURATION MODIFICATION

One of the earlier attempts at artificially manipulating the speaking rate was a scheme described by Fairbanks *et al.* [52]. They achieved an increase in speaking rate (time compression) by discarding intervals of speech of approximately 10-30 ms duration. A decrease in speaking rate (time expansion) was achieved by repeating 10-30 ms segments of speech. For normal-hearing listeners trained on a 50 word vocabulary, it was shown that time compression and expansion factors of upto 4 did not adversely affect intelligibility. However, a similar study [53] using elderly hearing-impaired listeners showed a decrease in intelligibility scores for both time compression and time expansion. The Picheny *et al.* study [47] too reported a similar deterioration in intelligibility for hearing-impaired listeners for both time compression and time expansion.

Lengthening consonant duration, using words or nonsense syllables, has yielded mixed results on speech recognition by hearing-impaired listeners. Gordon-Salant [11] reported no performance improvement by young and elderly normal-hearing listeners for consonants doubled in duration relative to their natural lengths. Furthermore, when duration lengthening was combined with consonant amplification, significantly more consonant confusions occurred than for consonant amplification alone. These results were observed for stimulus presentation at both 75 and 95 dB SPL. Similar results were observed for elderly listeners with mild to moderate sloping losses [49]. Montgomery & Edge [48] reported a 5% improvement in recognition of consonants lengthened by 30 ms, with and without consonant amplification, for a group of hearing-impaired subjects listening to the stimuli at 95 dB SPL. However, in the same study, a second group of hearing-impaired subjects listening at 65 dB SPL demonstrated no benefits from lengthened consonants.

The indeterminate effect of consonant lengthening on the consonant recognition performance might be due to any number of differences between the above studies. These may include differences in stimuli, subject background, amount of consonant lengthening, effects of signal processing, etc. Another confounding factor is that the acoustic segments associated with consonant phonemes (e.g., voice onset time, stop gap, etc.) were found to increase in a nonuniform manner [9]. Furthermore, the average increases in duration varied somewhat with phoneme class.

It may be questioned whether universal lengthening of consonants would improve the overall recognition performance. One predictable phoneme error from expanded consonant durations could occur between unvoiced stops and fricatives. Unvoiced consonants with lengthened bursts have been confused with fricatives [11]. Unvoiced stop bursts are known to have a relatively longer duration [54]. Hence, when the release bursts of voiced stops contain no periodicity, lengthening these bursts could elicit a unvoiced percept. Similarly, voiced stops with lengthened formant transitions could be confused with glides.

Revoile *et al.* [19] studied voicing distinctions for word-final fricatives by severely and profoundly hearing-impaired listeners. This consonant distinction can be cued by the duration differences in the vowels preceding voiced versus unvoiced word-final consonants. For vowels that were similar in duration in spoken consonant-vowel-consonant syllables, those preceding voiced fricatives were lengthened and those preceding unvoiced fricatives were shortened. Compared to the unaltered stimuli, significantly better perception of fricative voicing was shown by the listeners who relied on the vowel duration cue for consonant voicing decisions. However, the enhancement of vowel duration had no effect on distinctions among fricatives according to place of articulation. It thus appears that it would be more effective to apply temporal enhancements only to consonant distinctions that are cued by duration differences.

3.5 MOTIVATION BEHIND THE PRESENT INVESTIGATION

The studies by Picheny *et al.* [46, 47] have identified increased consonant amplitude (improved consonant/vowel intensity ratio) and increased consonant duration as two characteristics of highly intelligible speech. All the studies reported above have shown that listeners with mild to moderate hearing impairment stand to benefit from consonant amplification. Lengthening consonant duration has, however, yielded equivocal effects on speech recognition. As noted earlier, lengthened consonant duration is accompanied by nonuniform increases in the acoustic segments associated with the consonant phonemes. For example, in the case of stop consonants, these acoustic segments include the closure duration, the burst duration, the formant transition duration, and the voice onset time. Thus, merely increasing the consonant duration by replicating small segments within the consonant [11, 48] cannot be expected to reflect the changes associated with the acoustic segments. What is called for is some method whereby the effects, on recognition, of increasing the duration of each acoustic segment independently can be monitored. This can be achieved using a speech synthesizer.

Speech synthesis is an attractive tool for many reasons. One reason is that signals can be created in which the spectral, temporal, and intensity characteristics vary independently. This, in principle, allows one to separate the relative contribution of the various signal parameters to overall intelligibility. A second attraction of synthetic speech signals is their malleability to a continuum — a set of signals which differ along a single physical dimension or a small set of physical dimensions.

Segmentation of individual phonemes in natural speech stimuli is a major problem because coarticulatory effects create an overlap of the acoustic signals specifying consonant and vowel information. This problem is less severe in the case of synthesized stimuli where the various parameters are user defined and the boundaries are more readily identified.

A study by Van Tasell *et al.* [55] using hearing-impaired listeners has compared stop consonant identification performance in a synthesized continuum with that obtained for naturally produced syllables and obtained similar results for both the stimuli. This would suggest that data from synthesized continua have relevance with regard to performance with naturally produced speech.

Keeping in view all the above factors it was decided to investigate the effects of CVR modification and CD modification on synthetic speech stimuli. A modified form of the Klatt synthesizer [56] was used to obtain the stimuli. Though other classes of sounds like glides, liquids, diphthongs, affricates, and certain fricatives have been successfully synthesized, it was decided to limit the stimuli for the present investigation to the English stop consonants /p, t, k, b, d, g/. This was done in order to keep the test set as well as the feature contrasts (e.g., place, voicing) simple.

The above consonants used for synthesis are common to all regional accents in Indian English. The accompanying vowels were the cardinal vowels /ɑ, i, u/. Consonant-vowel and vowel-consonant syllables were used and prosodic features like stress and intonation do not play a significant role in the perception of these syllables. Hence, the results are not likely to be affected by individual perception of subjects on the basis of regional accents in Indian English.

The next chapter describes the properties of English stop consonants and the method adopted for synthesizing them.

CHAPTER 4
STOP CONSONANTS — THEIR PROPERTIES AND THEIR SYNTHESIS

4.1 INTRODUCTION

Synthesized stop consonants in different vowel environments have been chosen for the present investigation for reasons cited in the previous chapter. The present chapter first describes the properties of stop consonants in order to better understand the nature of the stimuli used. This is followed by a brief description of speech synthesizers. The synthesis of stop consonants using the cascade pole-zero synthesizer is also discussed.

4.2 PROPERTIES OF STOP CONSONANTS

Speech sounds are generated by the vocal organs which include the lungs, the windpipe, the larynx (containing the vocal cords), the pharynx, the nose, and the mouth. The organs lying above the larynx constitute the vocal tract. The shape of the vocal tract can be changed by the movements of the tongue, the lips, and the jaw. In between the vocal cords there is a variable opening called the glottis. It can affect the airflow from the lungs by opening and closing quite rapidly in a periodic or quasi-periodic manner. This results in a periodic complex tone, whose spectrum contains energy at harmonics covering a wide range of frequencies. This spectrum is subsequently modified according to the shape of the vocal tract. The rate of vocal cord vibration is called the fundamental frequency (F_0) and the period of vibration ($T_0 = 1/F_0$) is called the pitch period. Speech sounds produced while the vocal cords are vibrating are said to be voiced (e.g., /b/ as in "bat"). For some sounds the vocal cords do not vibrate, but remain open. Such sounds are called unvoiced, and result from turbulence at a constriction of the vocal tract above the glottis (e.g., /s/ as in "sat").

There are various types of sounds produced by a constriction in the vocal tract. Fricative consonants are characterized by a turbulent noise, and may consist of that noise alone (as in /f/), or may consist of that noise together with vocal cord vibration (as in /v/). Stops consonants involve a sudden rapid closing and opening of a complete constriction somewhere in the vocal tract. This stops the airflow briefly resulting in a reduction of acoustic energy, after which the airflow resumes and the energy increases. Stops too may be voiced (as in /b/) or unvoiced (as in /p/). Affricates may be treated as a combination of a stop and fricative. They involve both turbulence and a momentary obstruction to the airflow through the vocal tract (e.g., /tʃ/ as in "church"). Nasals, such as /m/ and /n/, are obtained by allowing air to flow through the nasal passages.

The above classes of sounds differ from each other in the manner in which they are produced. In addition, within each class, there are differences in the place of constriction (e.g., roof of the mouth, teeth, lips). The spectra of consonants are not static but vary as a function of time. However, vowels may have relatively stable spectra, at least for short periods of time.

Many speech sounds show peaks in the spectral energy at particular frequencies known as formant frequencies or formants. Formants are the resonances of the vocal tract [57, 58]. The formants are numbered, the one with the lowest frequency called first formant (F1), the next the second (F2), and so on. The formants depend upon the shape and dimensions of the vocal tract. Discrimination of most speech sounds is accomplished largely by the first two formants, but additional information is provided by the higher formants.

From the above discussion, it is clear that speech is a signal which varies in frequency, amplitude, and time. These variations can be displayed simultaneously by a device known as the speech spectrograph [59]. Fig. 4.1 shows the wideband spectrograms for the natural utterances */aka/* and */aga/* obtained using a digital spectrograph [60, 61]. The horizontal and vertical axes correspond to time and frequency respectively and the darkness of the pattern represents signal energy. Very dark areas indicate high concentration of energy at particular frequencies, while light areas indicate the absence of energy. The tracks corresponding to the various formants are clearly visible in the figure. Rapid changes are observed in the formants as the phoneme changes from the vowel to the stop and then from the stop to the ensuing vowel. These changes are known as formant transitions and they form important acoustic cues for consonant perception.

The plosive burst of energy accompanying the release of closure in the case of /k/ and /g/ are also observed in Fig. 4.1. The time from the start of the burst to the start of vocal cord periodicity is called the voice onset time (VOT) [62]. In the case of /g/, the vocal cord vibration (seen in the form of a voice bar) starts immediately after the burst. However, in the case of the unvoiced stop /k/, the VOT is seen to be larger.

The present investigation is limited to the stop consonants /p, t, k, b, d, g/. Nasal cavity does not play a role in the production of these consonants. The acoustic cues that contribute to the perception of a stop consonant can be summarized as follows: a closure or silence by occlusion of the articulators; a plosive burst of energy when the closure is released; formant transitions; and VOT [57, 58, 63, 64]. The following subsections discuss the importance of these acoustic cues in the identification of place of articulation

4.2.1 Burst Release

Liberman *et al.* [65] had concluded that stop consonants are "special" in that they can be identified as speech sounds only when they are presented in the context of a steady-state vowel. However, Stevens & Blumstein [2] observed that the gross shape of the spectrum at consonant release uniquely specified consonant place, irrespective of vowel context. The shape of the spectrum is related to the acoustic properties of the consonant burst as well as to the formant frequencies at consonant release, but not to the latter formant transitions into the following vowel. Subsequently, they (Blumstein & Stevens [66]) showed that synthetic consonant-vowel (CV) stimuli consisting only of burst plus 10 ms of voicing could be identified accurately in terms of place of consonant articulation.

Kewley-Port [67] suggested that the gross shape of the onset spectrum must be an insufficient cue for stop consonant place, because it does not incorporate the dynamic changes in spectral shape that occur during consonant release and subsequent articulatory movement toward the following vowel. She proposed three dynamic spectral features as invariant cues for stop consonant place, in which the tilt of the spectrum at burst onset cues bilabial place if it is falling, and alveolar place if it is flat or rising. Late onset of low-frequency energy functions as a cue for velar place, since velar stops are characterized by longer VOT than are bilabials or alveolars. Finally, the presence of mid-frequency peaks extending over time is a cue for velar place; these peaks reflect the resonant characteristics of the cavity anterior to the velar constriction. All of these features may be evaluated during the first 20-40 ms of the CV waveform.

Hence, regardless of the actual nature of the invariant cue for consonant place, the data of both Stevens & Blumstein [2] and Kewley-Port [67] indicate that it apparently resides in the first 20-40 ms of the syllable waveform.

4.2.2 Formant Transitions

Spectrographic analyses have revealed that formant transitions do not demonstrate acoustic invariance. Both the frequency location and direction of the transition vary as a function of vowel environment [68]. Nonetheless, many perceptual studies have demonstrated the importance of the second and third formant transitions in place identification for stops [1, 68, 69]. The studies by Blumstein & Stevens [66] have shown that while the entire formant transition may be a sufficient cue to place of articulation, it is not a necessary one.

4.2.3 Voice Onset Time

The phonemes /b, d, g/ are said to be voiced in contrast to the phonemes /p, t, k/, which are unvoiced. In general, the terms voiced and unvoiced imply either the presence or absence of vocal cord vibration during the articulation of the sound [62, 70]. However, for stops the situation is a bit more complex. The closure portion of a stop renders the speech silent and hence there is a gap in the acoustic pattern for both voiced and unvoiced stops. But if the vocal cord vibration is present during closure then a simple voice bar with energy confined to the first few harmonics (for voiced stops) appears at the base line (at low frequency) of the spectrogram.

When the vocal tract occlusion is released, turbulent noise generation (frication) continues for 10-40 ms exciting high-frequency resonances, before the vocal tract moves towards a position for the ensuing vowel. Unvoiced stops generally have larger duration of frication than the voiced ones (35 vs 20 ms) [71]. In voiced stops, vocal cord vibration either continues throughout the entire stop or start immediately after the burst. In the case of a syllable with an unvoiced initial stop, say /p/, the VOT is of the order of 50-70 ms.

Other important cues to initial-stop voicing are (a) a relatively low pitch (F_0) at vowel onset for voiced consonants; and (b) the presence of aspiration in the VOT interval of unvoiced stops [72].

4.3 SPEECH SYNTHESIZER

Speech synthesis is the process of producing an acoustic signal by controlling the model for speech production with an appropriate set of parameters. One of the first electrical synthesizers which attempted to produce connected speech was the "voder", reported by Dudley in 1939 [57], which followed the principle of separation of the excitation source and vocal tract. The advent of digital hardware and computers has revolutionized the development of speech synthesis. A software-based programmable speech synthesizer provides the flexibility for controlling the synthesis parameters as needed for generation of the test stimuli in psychoacoustic and speech perception studies. In these applications, speed requirement is not very critical and stimuli can be synthesized off-line. Out of the various types of synthesizers [58, 62], the formant synthesizer is particularly suitable for these studies, because synthesis parameters are directly related to perception features of the speech sounds.

Fig. 4.2 illustrates a simplified speech production model using the formant synthesizer. It consists of an impulse train generator exciting a cascade of resonators, simulating spectral shaping by the vocal tract. The resonator parameters, namely formant frequencies F_i and formant bandwidths B_i, are under user control so that different kinds of spectral shaping may be simulated. Unvoiced speech is synthesized by passing scaled random noise through a system that consists of a complex pole and a complex zero.

Alternatively, the resonators can be combined in parallel with their outputs summed together as shown in Fig. 4.3. In this case, the gain of each resonator should be selected carefully because zeroes are introduced in the transfer function in addition to the poles.

In a software-based formant synthesizer, developed by Klatt [56], the representation of the vocal tract transfer function is achieved by an all pole-model. The synthesizer uses 39 parameters that go into determining the output and as many as 20 of these can be varied as a function of time. Klatt synthesizer has the flexibility of utilizing either a parallel or a cascade structure. It can be used in two ways, either all-parallel mode or cascade/parallel mode as shown in Fig. 4.4.

As a modification to cascade/parallel all pole model Klatt synthesizer, a cascade pole-zero synthesizer has been developed at IIT Bombay [73, 74, 75] In this model the individual poles and zeroes are used to simulate the transmission in vocal tract by a cascade connection of resonators and antiresonators. This scheme is depicted in Fig. 4.5. Voicing is simulated by an impulse train of fundamental pitch period which is directed through the resonator and antiresonator (RGP and RGZ). The amplitude of voicing can be controlled. There are six resonators (R1 - R6) connected in cascade and five antiresonators (RZ1 - RZ5) connected in cascade. These antiresonators simulate zeroes in segments involving frication. Vowels are simulated by using voiced excitation source of resonators in cascade. Unvoiced fricatives and burst portions of unvoiced stops are simulated using frication excitation. For voiced stops and voiced fricatives (mixed mode excitation), the periodic impulse generator modulates the random noise generator. Nasalization is achieved using an additional resonator and an antiresonator, both of which are kept at some frequency (270 Hz for adult male) during non-nasalized sounds and the nasal zero frequency is increased during nasalization. For exact cancellation in non-nasalized utterances the frequencies as well as the bandwidths of the nasal poles and the nasal zeroes must be the same.

Transfer function zeroes are accounted for in the Klatt synthesizer by appropriate settings of the formant amplitude control of resonators in parallel configuration. Klatt [56] derived this data from trial-and-error attempts to match natural frication spectra. Parameter data for cascade pole-zero synthesizer, i.e., zero frequencies and bandwidths, were obtained from Klatt's amplitude control data using a program ZEROAN [74].

4.4 SYNTHESIS OF STOP CONSONANTS

The first step in synthesis is to generate parameter tracks, i.e., parameter variations as a function of time. These parameter tracks are required by the cascade pole-zero synthesizer. The synthesizer will give a good result if the parameter tracks are precisely specified.

The generation of parameter tracks is a trial-and-error procedure but is based on some *a priori* knowledge of the parameters Approximate formant tracks may be obtained by evaluating a natural segment of the utterance on a wideband digital spectrogram. This would also give a rough idea of the amplitude and bandwidth tracks. A digital spectrogram program [60, 61] has been used for this purpose. Formant transition parameters for naturally produced voiced stop consonant-vowel syllables reported by Kewley-Port [67] could also be used as a rough guide. Target values of various parameters could be obtained from the Klatt [56] paper.

A program PARTRC [74] is used for graphically editing the parameter tracks. The cursor controls can be used to mark the points graphically. All the parameters that are variable can be edited in the graphical mode. Excitation amplitude controls AV, AF, AH, and AVS are used to adjust overall intensity contour and mixture of periodic voicing to aperiodic noise. Several iterations and adjustments have to be carried out to obtain perceptual distinctiveness and quality in the synthesized samples. The program PZSYNTH [74] takes the parametric track file, obtained using PARTRC, as its input and generates the synthesized speech file.

The parametric tracks for some of the stop consonant stimuli used in this study are given in Appendix B. If synthesis is done with a constant fundamental frequency F_0, the speech will sound "machine-like". Hence, as suggested by Klatt [56], a "natural" pitch contour has been added by specifying that F_0 falls off linearly or rises linearly over the interval depending upon the context. Average syllable duration was 300 ms in conformity with the duration of such syllables in average conversational speech [76].

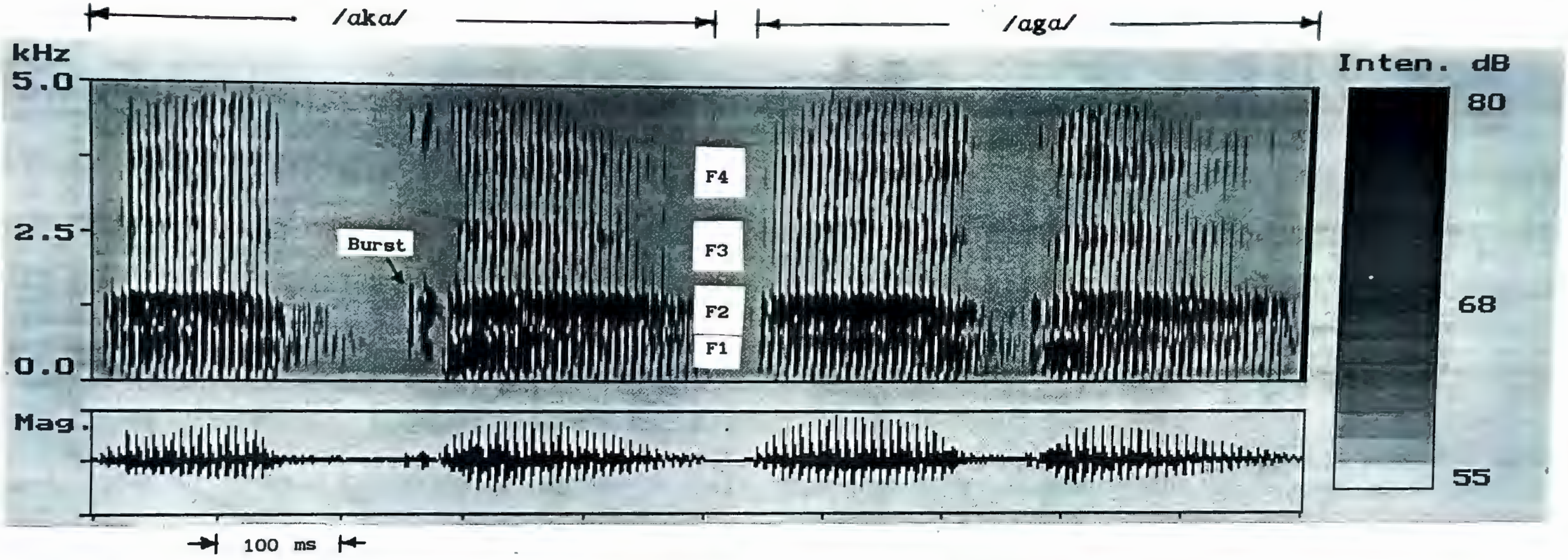


FIG. 4.1 Wideband spectrograms for/*aka*/ and /*aga*/.

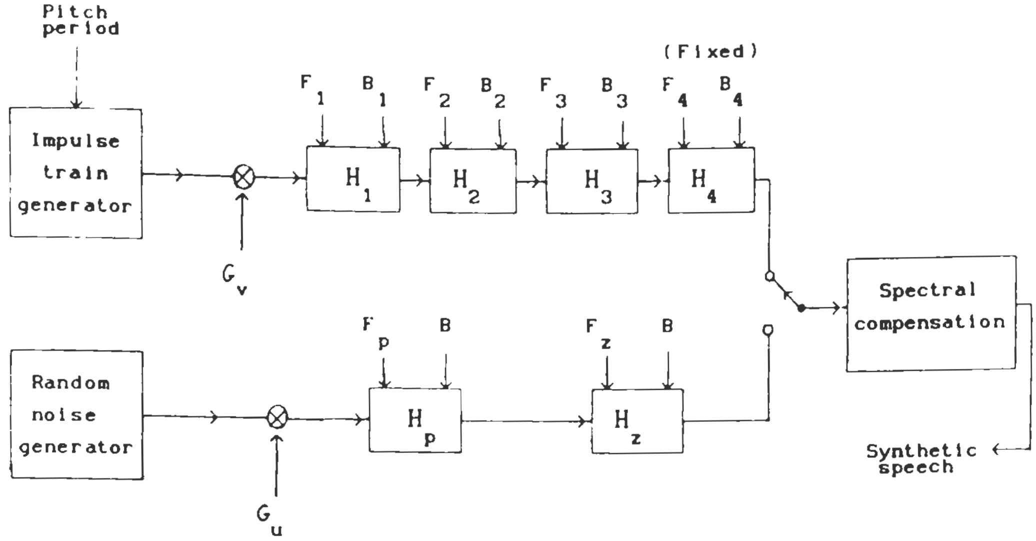


FIG. 4.2 Speech production model using a formant synthesizer Source: [77], p 103.

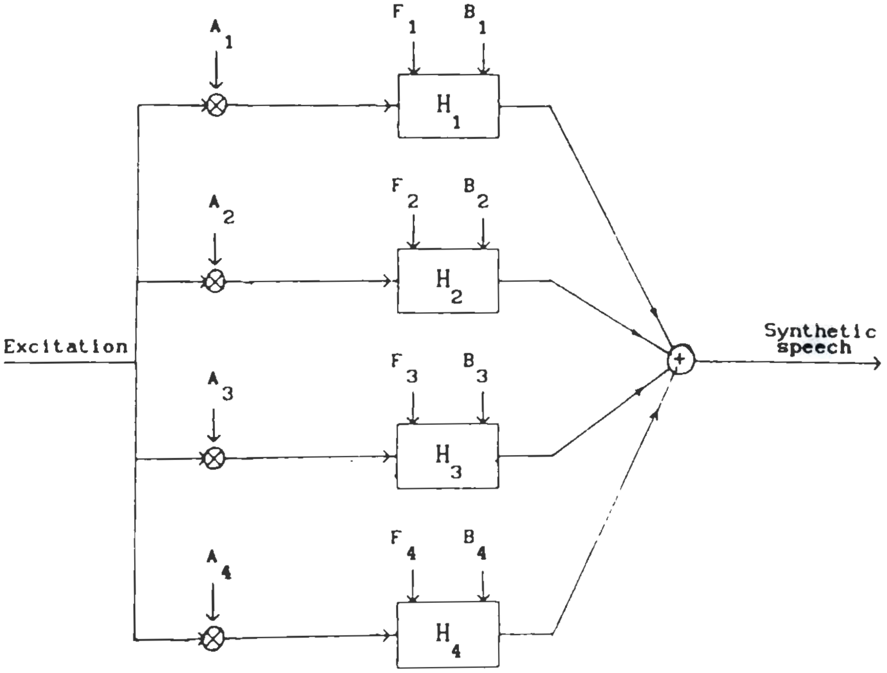
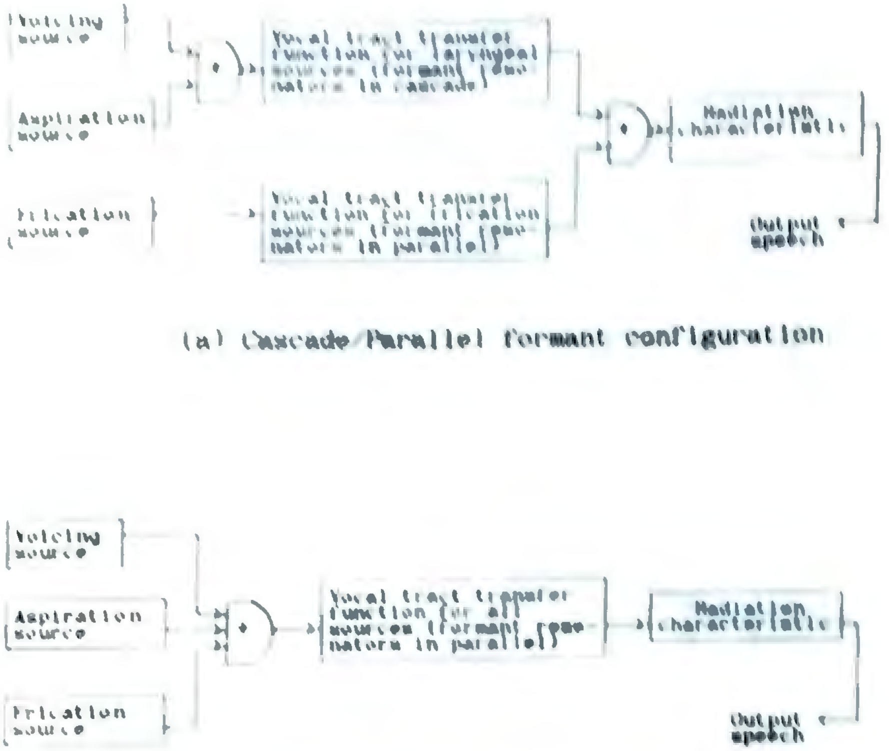


FIG. 4.3 Parallel combination of resonators of a formant synthesizer. Source: [77], p 105.


Voicing source
Aspiration source
Frication source
+
Vocal tract transfer function for laryngeal sources (formant resonators in cascade)
Vocal tract transfer function for frication sources (formant resonators in parallel)
+
Radiation characteristic
Output speech
Voicing source
Aspiration source
Frication source
+
Vocal tract transfer function for all sources (formant resonators in parallel)
Radiation characteristic
Output speech

(a) Cascade/Parallel formant configuration

(b) Special purpose all parallel formant configuration

FIG 4.4 Klatt's synthesizers Source [56]

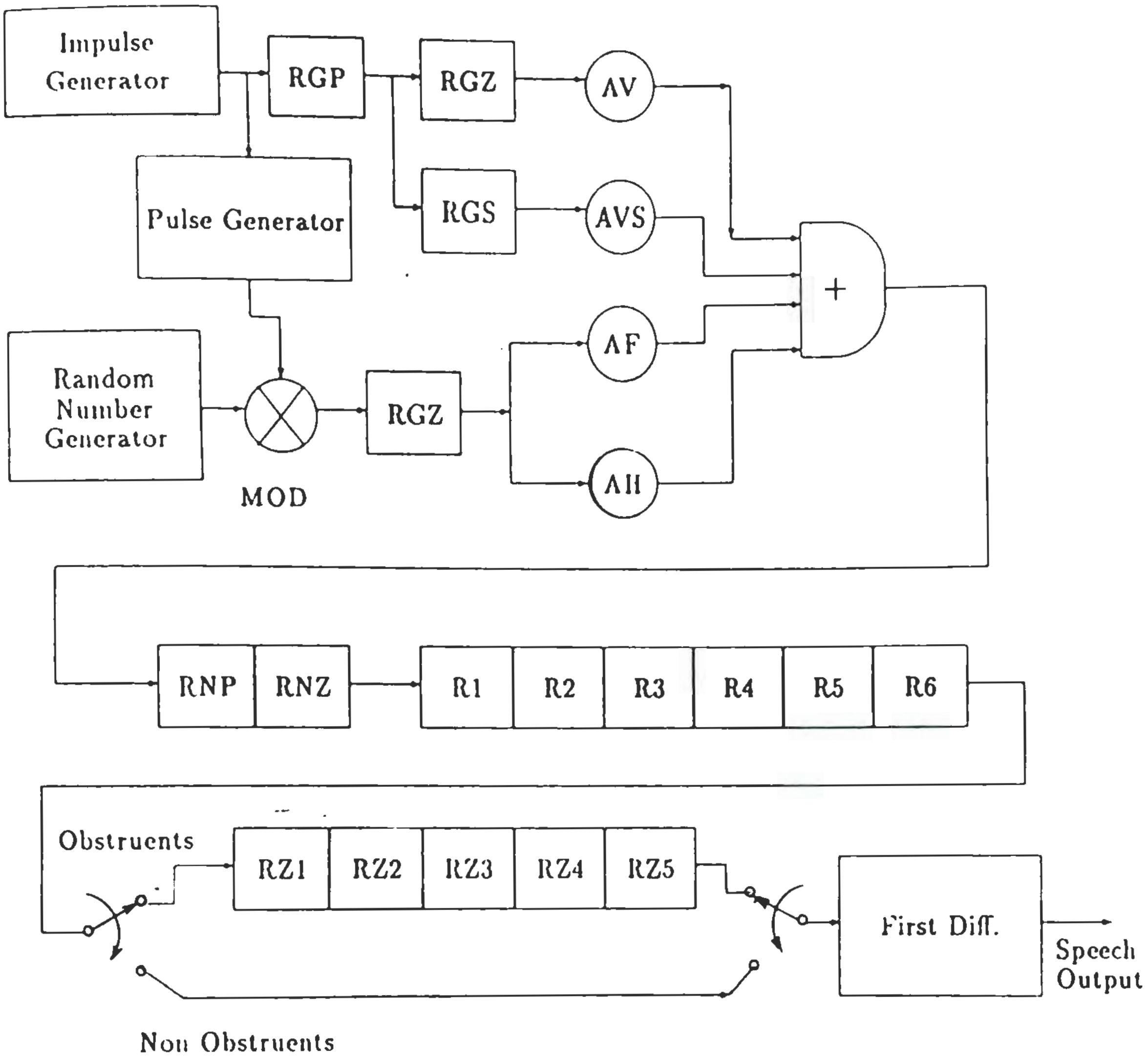


FIG. 4.5 Cascade pole-zero synthesizer. Source: [73]

CHAPTER 5
EXPERIMENTAL METHODOLOGY AND PERFORMANCE EVALUATION

5.1 INTRODUCTION

Speech discrimination test results are usually summarized by a percentage correct response score for many experimental runs. However, for a more detailed study of the received speech information, the results of each run are presented in the form of a stimulus-response confusion matrix, with rows corresponding to the stimuli and columns corresponding to the responses.

In this chapter, the methodology adopted for the experiments described in Chapters 6 and 7 is first discussed. Next, some of the techniques for analyzing and summarizing the confusion matrix data will be reviewed and the choice of the techniques used in the analysis of test results will be discussed.

5.2 EXPERIMENTAL METHODOLOGY

The purpose of the experiments was to evaluate the effect of certain types of speech processing on the perception of stop consonants by normal hearing subjects with simulated hearing impairment. The job of the subject was to listen to and identify the sounds presented to him over a headphone. Due to the repetitive nature of the experiments, a computerized test administration system was developed in order to automate the process. The details of the apparatus used as well as the test presentation procedure are discussed below.

5.2.1 Apparatus

The experiments measured the stimulus-response confusions for different stimulus sets. Experiments were carried out using a computerized test administration system [78] as shown in Fig. 5.1. The system was developed for an IBM-PC. Peripherals of particular importance included a terminal connected to the asynchronous serial port (RS-232), and a PC-based data acquisition card PCL-208 (from Dynalog Micro Systems, Bombay). The terminal was used for displaying the response choices on its screen and for obtaining subject responses from its keyboard.

For presentation, the data files were played back at a rate of 10 k samples/sec through the D/A port of the data acquisition card. The D/A output was passed through a 7th-order active elliptic low-pass filter [78], and a power amplifier. The stimuli were presented to the right ear through a pair of audiometric headphones DR-59 (Eliga, Japan).

Further details of the hardware and software for the experimental control are provided in Appendix C.

5.2.2 Presentation Level and Simulation of Hearing Loss

Several studies have examined the effect of presentation level on stop consonant place identification in normal-hearing listeners [79, 80, 81]. The results indicate that identification performance deteriorates at either very high (>90 dB SPL) or very low (<35 dB SPL) presentation levels. Accordingly, for the experiments described here, a presentation level of 75 dB SPL has been chosen throughout, which also happened to be close to the most comfortable listening level for the subjects.

In order to simulate hearing impairment in normal-hearing subjects, a standard practice is to present the stimuli in a background of noise. Some studies have employed multi-talker babble as the background noise [11]. However, due to its non-stationary character, the effective masking it may provide during syllable stimulus presentation is unpredictable. Alternatively, stationary noise filtered to have speech waveform spectrum [3] has been used to simulate hearing impairment in normal-hearing subjects. However, filtered signals presented to normal-hearing subjects provide a better simulation of hearing loss associated with conductive pathology than with sensorineural impairment [82].

Some investigators have studied speech-recognition performance of normal-hearing subjects under conditions in which their thresholds have been elevated by the addition of broadband noise [83]. The masking process responsible for threshold elevation is believed to be predominantly of cochlear origin [84]. In addition to simulating reduced dynamic range, broadband noise also more closely approximates the loudness-growth function of listeners with sensorineural hearing loss [85]. Other studies have evaluated frequency selectivity [86], and minimum detectable gap duration [87]. The results of all these studies suggest that presentation of broadband noise to normal-hearing listeners may be a valid method of simulating threshold elevation and its consequences.

Based on the studies reported above for masking broadband noise, it was decided to employ broadband noise in the present investigation. This noise was synthesized using the frication noise source in the synthesizer program and adjusting the parameters AF and the pole-zero locations to obtain a fairly uniform noise spectrum within the range of speech frequencies.

5.2.3 Presentation Procedure

Each experimental run consists of a number of presentations of the stimuli, in a randomized order with certain uniformity constraints (given in Appendix C). For each presentation, the response choices are displayed on the subject screen. Each choice corresponds to a key on his keyboard. Before each presentation, a "listen" message is flashed on the screen to alert the subject. The subject selects the response (guesses if necessary) by hitting the appropriate key. In order to minimize any bias in responses, the order of items in the response list and the positions of the correct response are also uniformly randomized. The response and the response time are recorded. At the end of each run, the stimulus-response confusion matrix, the recognition score (defined in Section 5.3), and the mean response time are stored. The experimental run can be administered with a feedback to the subject indicating the correct responses or without feedback.

In a session, the subject first listens to each stimulus item separately and any number of times. After becoming familiar with the stimuli, he then proceeds to an experimental run with feedback followed by runs without feedback. At the request of the subject, runs with feedback may be administered between the runs without feedback While all the data are recorded, only the no-feedback matrices were used in the analysis.

5.3 PERFORMANCE EVALUATION

The stimulus-response confusion matrices are given either as the stimulus-response frequencies or the probabilities estimated from the measured frequencies [12].

Let x be the set of n stimuli $\{x_1, x_2, \ldots, x_n\}$ and y be the set of n responses $\{y_1, y_2, \ldots, y_n\}$. If $N(x_i)$, $N(y_j)$, and $N(x_i;y_j)$ are the frequencies of stimulus x_i, response y_j, and the stimulus-response pair $(x_i;y_j)$ in a sample of N observations, then the probabilities can be estimated as follows:

$$\begin{aligned} p(x_i;y_j) &= N(x_i;y_j)/N, \\ p(x_i) &= N(x_i)/N = \sum_{j=1,n} p(x_i;y_j) \\ p(y_j) &= N(y_j)/N = \sum_{i=1,n} p(x_i;y_j) \end{aligned} \qquad (5.1)$$

The diagonal cell entries $(i=j)$ correspond to correct responses and the off-diagonal entries $(i\neq j)$ correspond to errors. It is usually difficult to study the error patterns in the confusion matrix, particularly if the matrix size is large. Hence, there exists a need to reorganize the data in forms which are easier to comprehend. Some of the techniques for analyzing and summarizing the data in the confusion matrices are discussed below.

5.3.1 Recognition Score

The sum of diagonal entries in a confusion matrix gives the probability of correct responses and is known as the recognition or articulation score R_s

$$R_s = \sum_{i=1,n} p(x_i;y_i) \tag{5.2}$$

Although this score is useful, it does not provide any information on the distribution of errors. One way to generalize the recognition score is to derive a smaller confusion matrix by combining the stimuli and responses into groups, in such a way that confusions within the groups are more likely than those between the groups [12]. If some of the stimuli share a common feature, then these stimuli can be grouped in accordance with this feature and the resulting new matrix will give the recognition score for the transmission of this feature. Thus a number of recognition scores can be used for specifying the transmission of various features.

5.3.2 Information Transmission Analysis

The recognition score has the disadvantage that it could be affected by a subjects' response bias. For example, if the subject adopted a strategy of simply giving the same phoneme response for all the presentations, the percent-correct score for that phoneme would be artificially high (chance scoring). Such a problem would not arise if we expressed the results in terms of the relative information transmitted.

Information transmission analysis as used by Miller & Nicely [12] provides a measure of covariance between stimuli and responses. This method uses mean logarithmic probability (*MLP*) measure of information [88, 89].

The information measures of the input stimulus x and output response y, $I_s(x)$ and $I_r(y)$ respectively, are given by the following functions:

$$I_s(x) = MLP(x) = -\sum_i p(x_i)\ log_2[p(x_i)] \text{ bits,} \tag{5.3}$$

$$I_r(y) = MLP(y) = -\sum_j p(y_j)\ log_2[p(y_j)] \text{ bits} \tag{5.4}$$

An *MLP* measure of covariance of stimulus-response is

$$\begin{aligned} I(x;y) &= MLP(x) + MLP(y) - MLP(xy) \\ &= -\sum_{i,j} p(x_i;y_j).\log_2 \frac{p(x_i)\ p(y_j)}{p(x_i;y_j)} \end{aligned} \tag{5.5}$$

The relative transmission from x to y is given by

$$I_{tr}(x;y) = I(x;y)/I_s(x) \tag{5.6}$$

Since $I_s(x) \geq I(x;y) \geq 0;\ 1 \geq I_{tr}(x;y) \geq 0.$

It has been observed by Miller & Nicely [12] that, like most maximum likelihood estimates, this estimate will be biased to overestimate $I(x;y)$ for small samples. If the sample size is large enough then the bias can be safely ignored.

The above measure of information transmission takes into account the patterning of errors and the score that can be obtained by chance alone. Fig. 5.2 [78] shows the relationship between recognition score R_s and relative information transmitted I_{tr} for a special case when the stimuli have equal probabilities, correct responses are equally distributed among the diagonal cells, and response errors are equally distributed among the off-diagonal cells. That is, if the recognition score is R_s and the number of stimuli (and responses) is n, the cell entries are:

$$p(x_i;y_j) = (R_s)/n \qquad i=j$$

$$= \frac{1-R_s}{n^2-n}, \qquad i\neq j$$

$$p(x_i) = p(y_j) = 1/n \qquad (5.7)$$

It is to be noted that when R_s is at chance value, $I_{tr}(x;y)$ is zero. Thus we have a scoring system in which a score of 0% information transmitted indicates chance identification of the phoneme and a 100% score represents perfect identification.

The relative information transmission measure can also be applied to the matrices derived from the original confusion matrix by grouping the stimuli in accordance with certain features. The relative importance of these features may then be evaluated. Alternatively, the transmission performance can be measured in the context of specific features.

While recognition scores are easiest to compute and interpret, information transmission analysis has the merit that it measures the covariance between the stimuli and responses and hence takes into account the relatedness of the two. A number of studies have adopted the information transmission analysis approach to interpret their results [78, 90, 91, 92, 93].

5.3.3 Response Time

In addition to recognition scores and information transmission analysis, average response times are also considered as another possible measure of comparing the test stimuli processed differently. A decrease in the average response time by the subject for a particular experimental condition A over that for condition B may be another way of asserting the superiority of the processing used for the stimuli in condition A.

5.4 DISCUSSION

This chapter first described the methodology adopted for the experiments described in the following chapters. Next, some of the different methods of analyzing and summarizing the data in stimulus-response confusion matrices were considered. While recognition scores are easiest to calculate and interpret, information transmission analysis has the advantage that it measures the covariance between the stimuli and responses and hence takes into account the relatedness of the two. In addition to these two measures, average response time was also considered as a possible measure of comparing the test stimuli processed differently.

A program was written for computing recognition scores and carrying out information transmission analysis. Brief descriptions of the program, analysis method, and an example of the analysis are provided in Appendix D. The effect of sample size on relative information transmitted is considered in Appendix E.

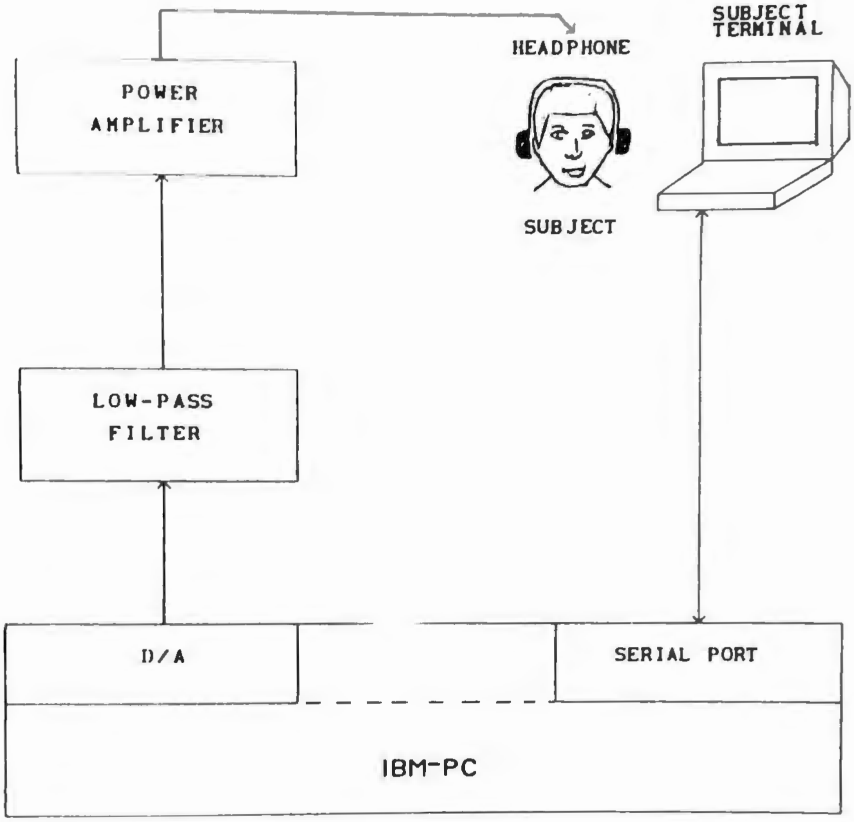


FIG. 5.1 A block diagram of the experimental set-up used in the computerized test administration system.

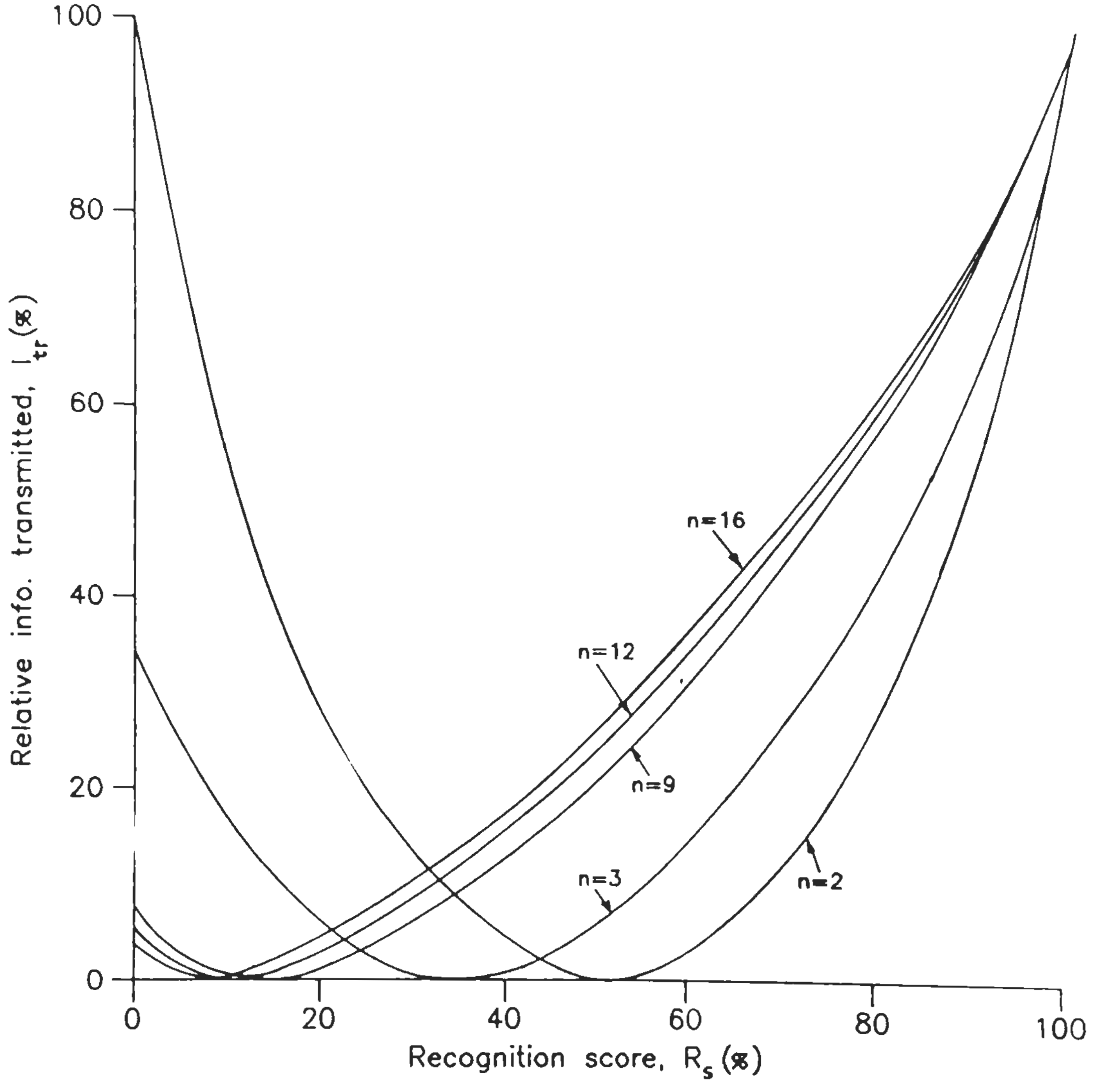


FIG. 5.2 Recognition score versus relative information transmitted for a special case when the correct responses are equally distributed among the diagonal entries and the errors are equally distributed among the off-diagonal entries in the stimulus-response confusion matrix (Eqn. 5.7). n = number of items. Source: [78]

CHAPTER 6

EFFECT OF CONSONANT-TO-VOWEL INTENSITY RATIO MODIFICATION

6.1 INTRODUCTION

It was seen in Chapter 4 that increased consonant-to-vowel (C/V) intensity ratio was one of the characteristics of clear (intelligible) speech. In the present chapter, the results of experiments performed to study the effects of consonant-to-vowel intensity ratio (CVR) modification on the recognition of stop consonants are presented.

The results and implications of varying CVR on the recognition of stop consonants in the CV and VC contexts of various vowels are presented. The effect of vowel context on consonant recognition as also impairments in vowel recognition, if any, with increasing CVR are studied. In addition to comparing the effect of CVR for different sets of stimuli on the basis of percentage consonant recognition scores, information transmission analysis has been used to study the effect of CVR on the transmission of features like place and voicing. The effect of CVR on the response time of the subjects is also considered.

6.2 TEST STIMULI

Nonsense syllables were used as stimuli instead of words in order to maximize the contribution of acoustic factors to confusions and to minimize the contribution of linguistic factors [92]. In a computerized test administration system of the type used here (described in Chapter 5), there is a limit to the number of stimuli that can be displayed on the subject terminal as possible choices to the presented stimulus. It is preferrable to limit the number of stimuli to a maximum of 12 to 16 in closed-set experiments.

The aim of the experiments was to study the effects of increasing CVR on consonant recognition in terms of the following: vowel context, vowel impairments, consonant position in the syllable, and response times. In addition, the effect of CVR on the transmission of information with respect to vowel context and features of place and voicing were also to be studied. As indicated in Chapter 3, the stop consonants /p, t, k, b, d, g/ in the CV and VC contexts of the vowels /ɑ, i, u/ have been chosen as stimuli for the present investigation. In order to limit the number of stimuli in each test, two sets of stimuli have been used as given below:

1) Unvoiced consonants /p, t, k/ with the cardinal vowels /ɑ, i, u/ yielding nine CV and nine VC syllables. The CV syllables were /pɑ, tɑ, kɑ, pi, ti, ki, pu, tu, ku/ while the VC syllables were /ɑp, ɑt, ɑk, ip, it, ik, up, ut, uk/. The tests performed with these stimuli are identified in the subsequent discussion as the CV9 and VC9 tests respectively. Different vowel environments have been chosen in order to study how consonant recognition was affected by vowel context with increasing CVRs. In addition, it was to be seen whether vowel recognition was impaired with increasing CVRs. The effect of increasing CVR on the overall

information transmission as well as transmission of information with respect to consonant and vowel features were also to be studied.

2) Consonants /p, t, k, b, d, g/ with vowel /ɑ/ yielding six CV syllables /pɑ, tɑ, kɑ, bɑ, dɑ, gɑ/ and six VC syllables /ɑp, ɑt, ɑk, ɑb, ɑd, ɑg/. The tests performed with these stimuli are identified in the subsequent discussion as the CV6 and the VC6 tests respectively. These experiments were performed to study the effect of varying CVR on consonant recognition when the stimuli included both voiced and unvoiced stops. The effect of CVR on the overall information transmission as well as transmission of information with respect to place and voicing features were also to be studied.

The digitized waveform of the synthesized stimulus was first displayed on the PC monitor using a program written for the purpose. The user could manipulate cursors to isolate the time waveform segments to be processed or reproduced aurally via the D/A converter. Using this interactive system, the consonant and vowel sections were defined for each syllable. However, the terms consonant and vowel should be used with caution. Repp [94] and others have commented that "consonant" and "vowel" are linguistic categories that are the end results of complex perceptual and cognitive processes. Here, both these terms are used to designate the relevant acoustic properties pertaining to them. The consonant and vowel segments were identified after repeated visual and auditory monitoring.

Once the segments were identified, the C/V intensity ratio for each digitized syllable was determined by calculating the mean of the squared amplitudes (average power) of the sampled points within the consonant and vowel segments and then taking the ratio between them. C/V intensity ratio was then expressed in dB. Calculations for the vowels were based on only the initial 100 ms of the vowel in the CV context and on the final 100 ms of the vowel in the VC context. Table 6.1 gives the durations of the consonantal segments and the C/V intensity ratios for the synthesized stimuli in both the CV and VC contexts. It is observed that if an attempt is made to increase the C/V intensity ratio of these stimuli, the consonants in certain syllables like /b*a*/, /*a*p/, /up/, and /*a*d/ are the first to become more intense than their adjoining vowels. Since the experiments reported here have been performed with closed sets of limited stimuli, it was felt appropriate to prevent the "weak-vowel" cue from resulting in higher recognition scores for these syllables. Hence, the CVR modification was limited to a maximum of 12 dB.

Treating the synthesized syllables referred to in Table 6.1 as the most "natural" representatives, four new versions of each syllable were synthesized by modifying the C/V intensity ratio by +3, +6, +9, and +12 dB. The durations of the consonant and vowel segments were not altered. Thus each syllable had five versions, one with no modification and the other four with CVR modification of +3, +6, +9, and +12 dB. To simulate hearing-impairment in normal-hearing listeners, each stimulus was mixed with synthesized broadband noise under three SNR conditions: no masking noise, and masking noise with 12 dB SNR and 6 dB SNR. These noise conditions were obtained by using a program SNR which scaled the synthesized noise data keeping the level of the speech signal fixed. Thus the CV9 test included 9 CV syllables under 15 conditions (5 CVR modifications X 3 SNRs), with a total of 135 test stimuli. The VC9 test too involved 135 stimuli. The CV6 and VC6 tests included 6 syllables under 15 conditions with a total of 90 test stimuli each. Appendix B includes sample spectrograms for /kɑ/ for different CVR modifications.

6.3 EXPERIMENTAL METHOD

Five normal-hearing subjects in the age-group of 21 to 35 years participated in the experiments. All the subjects had pure-tone auditory thresholds within 20 dB of the normal hearing standards. The test ear was the right ear for all the subjects

The details of the apparatus and the presentation procedure have been discussed in the previous chapter. There were 15 experimental conditions (5 CVR modifications X 3 SNRs) for each of the four tests. For each subject, the CV9 test for each experimental condition was carried out in about 6-8 experimental runs, inclusive of at least one run with feedback. Each experimental run took at least 5-6 min for completion and included 5 randomized presentations of each of the 9 syllables thus resulting in 45 presentations per run. Thus the 3 SNR conditions together required about one-and-a-half to two hours for completion. The five CVR conditions required a total of about eight to ten hours for completion. The VC9 test too required a similar period of time by each subject for completion. The CV6 and

VC6 tests required about five to seven hours each for completion. Hence this experiment involved an involvement of about twenty five to thiry five hours per subject. Taking each subject's convenience into account, the tests were spread over a two-month interval for the five subjects tested.

For each subject, test runs with different sets of stimuli were randomized in order to reduce biases due to learning effects. At the end of each run, the subject responses were stored in the form of a stimulus-response confusion matrix. In addition, the overall recognition score, the average response time and its standard deviation were also stored. For each experimental condition, a number of runs (not necessarily successively) were carried out until the scores more-or-less stabilized. For the final analysis, three of the more stable confusion matrices (stable in terms of recognition scores) were combined together using the program CUMMAT for each subject for each experimental condition.

6.4 TEST RESULTS FOR CV9 AND VC9 STIMULI

The results for the CV9 and VC9 sets of stimuli are presented here. First we look at the recognition scores, followed by the results of information transmission analysis, and response time. Although the recognition score is useful, it could be affected by a subject's response bias. Furthermore, it does not provide any information on the distribution of errors. On the other hand, information transmission analysis takes into account the patterning of errors and is not affected by the subject's response bias. Response times were recorded as a possible measure of differences between test stimuli processed differently.

6.4.1 Recognition Scores

The consonant recognition scores obtained by the five subjects for the CV9 test are shown in Table 6.2. As the trend in scores under different conditions for individual subjects was similar, these scores have been averaged across the five subjects and are given in the last row of Table 6.2 and plotted in Fig. 6.1. In the no masking noise case, fairly uniform near-perfect scores are obtained for all the CVRs. For the 12 dB SNR case, the score increases from 62% (0 dB CVR modification) to 88% (12 dB CVR modification) — a total increase of 26%. The corresponding scores in the 6 dB SNR case are 52% (0 dB CVR modification) and 80% (12 dB CVR modification) — an increase of 28%. Thus, it is observed that increasing CVR does improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects [95].

In order to study the effect of vowel context on consonant recognition, three new confusion matrices were derived from the original confusion matrix in the context of the three vowels. The consonant recognition scores were obtained using Eqn. 5.1 and these are given in Table. 6.3. It is observed that the recognition scores in the /ɑ/ context for different CVRs are near-perfect for the no masking noise case. The corresponding scores in the /i/ and /u/ contexts are slightly lower. In general, there does not seem to be any detrimental effect on consonant recognition due to increasing CVR in all the vowel contexts. A small increase in recognition scores with increasing CVR is noted in the /u/ context.

For the 12 dB SNR case, the recognition score in the /ɑ/ context increases from 66% at 0 dB CVR modification (i.e., no modification) to 91% at 12 dB CVR modification — a total increase of 25%. The corresponding increase in scores in the /i/ and /u/ contexts are 23% and 31% respectively.

For the 6 dB SNR case, the recognition score in the /ɑ/ context increases by 32% as the CVR is increased by 12 dB. The corresponding increase in scores in the /i/ and /u/ contexts are 24% and 30% respectively.

The above results for the CV9 test show a similar pattern of score variation under different conditions in all the vowel contexts. However, in general, the corresponding scores are seen to decrease as we go from the /ɑ/ context to the /i/ context and then to the /u/ context. This finding is opposite to the results of Wang & Bilger [26] who obtained (for unmodified natural CV syllables) the highest scores for consonants paired with vowel /u/ and lower scores with vowel /ɑ/ for both quiet and noise. According to their study, consonants followed by vowel /i/ were the most difficult to identify. However, the results of the present study agree with those reported by Dubno & Levitt [28] for unmodified natural CV syllables and also the findings by Gordon-Salant [11] for stops (plosives) in natural CV syllables at 10 dB CVR modification.

For the CV9 test, the most frequent consonant confusions exceeding 20% for the no masking noise case are ku/pu (33%) at 0 dB CVR modification (the notation ku/pu implies /ku/ confused as /pu/); Other consonant confusions are ku/pu (23%) at 3 dB CVR modification; ku/pu (29%) at 6 dB CVR modification; pi/ti (25%) and ku/pu (22%) at 9 dB CVR modification; and no appreciable confusions at 12 dB CVR modification. For the 12 dB SNR case, they are ku/pu (44%), pi/ti (34%), pu/ku (28%), pa/ta (24%), ka/ta (24%), ti/ki (24%), and ku/tu (22%) at 0 dB CVR modification; ku/pu (62%) ti/ki (28%), pu/ku (27%), and tu/pu (23%) at 3 dB CVR modification; ku/pu (34%) pu/ku (24%) at 6 dB CVR modification; ku/pu (34%) at 9 dB CVR modification; and ku/pu (25%) at 12 dB CVR modification. For the 6 dB SNR case, the consonant confusions are ka/pa (40%), pi/ti (38%), pu/ku (30%), ku/pu (28%), ku/tu (28%), ti/pi (25%), ka/ta (24%), ti/ki (24%), tu/ku (24%), ta/ka (22%), at 0 dB CVR modification; ku/pu (48%), ta/ka (31%), ka/ta (30%), pu/ku (29%), tu/ku (29%), ki/pi (25%), pi/ti (24%), and pu/tu (23%) at 3 dB CVR modification; ta/ka (29%), ka/ta (25%), ku/tu (25%), pi/ti (23%), and pu/tu (23%) at 6 dB CVR modification; ku/pu (29%), pu/tu (24%), and ku/tu (22%) at 9 dB modification; and ku/pu (40%) at 12 dB CVR modification. A significant feature observed above is that for the simulated hearing impairment conditions, the number, as well as the severity, of consonant confusions decrease as the CVR is increased. Examination of the confusion matrices does not reveal any symmetry in these confusions. Furthermore, no appreciable across-vowel confusions are observed.

The consonant recognition scores obtained by the five subjects for the VC9 test are shown in Table 6.4. As the trend in scores under different conditions for individual subjects was similar, these scores have been averaged across the five subjects and are given in the last row of Table 6.4 and plotted in Fig. 6.2. The scores for all the five CVR modifications are near-perfect in the no masking noise case. For the 12 dB SNR case, the score increases from 85% (0 dB CVR modification) to 93% (12 dB CVR modification) — a total increase of 8%. The corresponding scores in the 6 dB SNR case are 73% (0 dB CVR modification) and 93% (12 dB CVR modification) — an increase of 20%. Thus, in the VC9 test too, increasing CVR is seen to improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects.

The consonant recognition scores in terms of vowel context for the VC9 test are given in Table 6.5. Perfect or near-perfect consonant recognition scores are obtained in the no masking noise case for the three vowel contexts for all the five CVR modifications. For the 12 dB SNR case, the recognition score in the /ɑ/ context increases from 85% at 0 dB CVR modification to 96% at 12 dB CVR modification — a total increase of 11%. The corresponding increase in scores in the /i/ and /u/ contexts are 8% and 18% respectively. For the 6 dB SNR case, corresponding increase in scores are 21% in the /ɑ/ context, 10% in the /i/ context, and 27% in the /u/ context.

Thus in the VC9 test too, the corresponding recognition scores under various conditions are seen to generally decrease from the /ɑ/ context to the /i/ context and then to the /u/

context However, the scores in the /i/ and /u/ contexts for CVR modification ≥ 6 dB in all the noise cases are much more closer to each other than they are in the CV context. For the 6 dB SNR case at 12 dB CVR modification, the score in the /u/ context is more than that in the /i/ context (92% versus 86%). It is, however, observed in general that inspite of modifying the consonant-vowel intensity ratio, these results agree with the results of Dubno & Levitt [28] who had reported (for unmodified natural VC syllables) the highest scores for syllables containing the vowel /ɑ/ and lower scores for syllables with the vowels /u/ and /i/.

An examination of the relevant confusion matrices for the VC9 test shows that there are no appreciable consonant confusions in both the no masking noise and 12 dB SNR conditions for all the CVRs. The consonant confusions involving place for the 6 dB SNR condition are ut/up (45%) and ɑt/ɑp (25%) at 0 dB CVR modification; and ɑt/ɑp (22%) at 3 dB CVR modification. There are no place confusions at higher CVRs. However, vowel confusion is observed at higher CVRs involving ip/up. This confusion is found to increase from ip/up (23%) at 6 dB CVR modification to ip/up (27%) at 9 dB CVR modification and ip/up (34%) at 12 dB CVR modification. It is noted that this across-vowel confusion is limited to the labial /p/ context. No other appreciable across-vowel confusions were observed. Thus, there appears to be a limit on the amount of CVR modification that may be used, beyond which certain vowel confusions may begin to surface. This limit, as well as the specific vowel confusions may again depend upon the test set involved

Thus, the results of both the CV9 and VC9 tests indicate that increasing CVR does improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects. A comparison of scores in Fig. 6.1 and 6.2 shows that the recognition scores for stops in the VC context are generally higher than those in the CV context for all the conditions considered here. Hence, this result is independent of consonant amplification.

Fig. 6.3 gives the pattern of confusions observed in the identification of place of articulation for different conditions in the CV9 and VC9 tests. For the no masking noise case in the CV9 test, there is no appreciable change in either the pattern of confusions or the overall level of confusions as the CVR is increased. For the 12 dB SNR case too, the pattern of confusions is more-or-less unaffected with increase in CVR. However, the overall confusions are seen to decrease with increase in CVR. For the 6 dB SNR case, the overall confusions are found to decrease noticeably for CVR modifications of 9 dB and above. However, the pattern of confusions remains largely unaffected for all the CVR modifications in the CV9 test.

For the no masking noise case in the VC9 test, near-cent percent recognition scores were obtained for all the five CVR modifications. For both the 12 dB and 6 dB SNR cases, the pattern of confusions does not change appreciably. However, the overall confusions reduce much more appreciably with increase in CVR than in the CV9 test. The confusions are almost reduced to the no masking noise level at 12 dB CVR modification. In particular, the labial/alveolar (LA/AL) and alveolar/labial (AL/LA) confusions decrease appreciably with increasing CVR. It is observed from Fig. 6.3 that increase in CVR is more effective in bringing down the overall level of confusions in the VC context as compared to the CV context. This suggests that increasing CVR suppresses forward masking of the consonant by the vowel more effectively than backward masking.

The consonant recognition scores obtained for the CV9 and VC9 tests were subjected to the paired t-test to examine the statistical significance or otherwise of the recognition scores under different levels of modification as compared to those with no modification. As seen from Table 6.6, no significant changes are seen when C/V ratio is increased in the no masking noise case for the CV9 test. However, for lower SNR conditions, the improvements in scores with increasing CVR are seen to be significant. For the 12 dB SNR case, significantly better results are obtained ($p<0.01$) for a CVR increase of 9 dB and above. For the 6 dB SNR case, the improvements are seen to be even more significant for 9 dB CVR modification ($p<0.005$) and 12 dB CVR modification ($p<0.001$).

Table 6.6 also includes the paired t-test results for the VC9 test. Here too, no significant changes are seen when the C/V ratio is modified in the no masking noise case. For the 12 dB SNR case, however, significantly better scores are obtained for CVR modifications of 9 dB ($p<0.001$) and 12 dB ($p<0.005$). A similar level of improvement is perceived at these CVR modifications for the 6 dB SNR case too.

Thus the results of the statistical analysis show that increasing CVR does improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects.

6.4.2 Information Transmission Analysis

Information transmission analysis results for overall information transmission and transmission of consonant and vowel features for the CV9 test stimuli, summarized in Table 6.7, show that the overall information transmission for the no masking noise case is fairly independent of the CVR modification. However, in the other two noise conditions there is an appreciable improvement in the overall information transmitted with increasing CVR; 59% at 0 dB CVR modification to 82% at 12 dB CVR modification for the 12 dB SNR case; 49% at 0 dB CVR modification to 73% at 12 dB CVR modification for the 6 dB SNR case. A similar trend is observed in the transmission of consonant information: 17% at 0 dB CVR modification to 62% at 12 dB CVR modification for the 12 dB SNR case; 8% at 0 dB CVR modification to 48% at 12 dB CVR modification for the 6 dB SNR case. The information transmitted about vowel is near-perfect for the no masking noise and 12 dB SNR conditions with a slight decrease for the 6 dB SNR condition.

The results for the information transmission of consonant and vowel features for the CV9 test are also plotted in Fig. 6.4. For the no masking noise case, the transmission of consonant feature is about 70% and does not get affected much by CVR modification. However, for SNRs of 12 and 6 dB, this feature is poorly transmitted for unmodified stimuli, and increases monotonically and appreciably with increasing CVR, up to 12 dB. CVR modifications exceeding 12 dB have not been considered due to the possibility of the "weak-vowel" cue mentioned earlier. The information transmission about vowel feature is higher than that for consonant feature for all the CVR modifications and SNRs considered.

Information transmission analysis results for the VC9 test stimuli, summarized in Table 6.8, show that the overall information transmission as well as transmission of consonant feature in the no masking noise condition is near-perfect for all the CVRs. As in the CV9 test, there is an appreciable improvement in the overall information transmitted with increasing CVR for the two noise conditions: 77% at 0 dB CVR modification to 97% at 12 dB CVR modification for the 12 dB SNR case; 66% at 0 dB CVR modification to 90% at 12 dB CVR modification for the 6 dB SNR case. A similar trend is observed in the transmission of consonant feature: 55% at 0 dB CVR modification to 97% at 12 dB CVR modification for the 12 dB SNR case; 40% at 0 dB CVR modification to 96% at 12 dB CVR modification for the 6 dB SNR case. The information transmission about vowel feature is near-perfect for the no masking noise and 12 dB SNR conditions with a slight decrease for the 6 dB SNR condition.

The results for the information transmission of consonant and vowel features for the VC9 test are also plotted in Fig. 6.5. For SNRs of 12 and 6 dB, the transmission of consonant feature is seen to increase appreciably with increasing CVR, and reach near-perfect levels at 12 dB CVR modification. The information transmission about vowel feature is, in general, higher than that for consonant feature for almost all the CVR modifications and SNRs considered. The only noticeable exception is at 6 dB SNR for which the information transmission of consonant is higher than that for vowel for CVR modifications exceeding about 10 dB.

6.4.3 Response Time

The response time for each presentation was recorded as a possible measure of differences between the test stimuli processed differently. The response times for the CV9 and VC9 tests, averaged across the five subjects, are summarized in Table 6.9.

For the no masking noise case in the CV9 test, CVR modification does not appear to have any appreciable effect on response time. For the two noise situations, there appears to be an improvement in subject response with increasing CVR. For the no masking noise case in the VC9 test, there appears to be no appreciable effect of CVR modification on response time. For the two noise cases, however, there appears to be an improvement in subject response times upto 6 dB CVR modification.

Thus, for both CV9 and VC9 tests, it is noted that the average response time certainly does not deteriorate and in most cases appear to indicate a quicker response by the subjects for increasing values of CVR.

The response times, averaged across the five subjects for the CV9 and VC9 tests, for various levels of modification were compared with those for no modification by subjecting the data to the paired t-test for statistical significance. The results are tabulated in Table 6.10. It is observed that, in general, there is no significant change in response time for all the conditions in both the tests. In one case, there is a significant improvement (decrease) in response time (p<0.1 at 12 dB CVRM for 6 dB SNR in the VC9 test).

6.5 TEST RESULTS FOR CV6 AND VC6 STIMULI

The results for the CV6 and VC6 stimuli are presented here. The percentage recognition scores are considered first, followed by information transmission analysis, and response times.

6.5.1 Recognition Scores

The consonant recognition scores obtained by the five subjects for the CV6 test are shown in Table 6.11. As the trend in scores under different conditions for individual subjects was similar, these scores have been averaged across the five subjects and are given in the last row of Table 6.11 and plotted in Fig. 6.6. The recognition scores for all the CVR modifications in the no masking noise case are seen to be near-perfect. For the 12 dB SNR presentation, the recognition score increases from 81% at 0 dB CVR modification to 90% at 3 dB CVR modification, then falls to 87% at 6 dB CVR modification and thereafter increases to 94% at 12 dB CVR modification. For the 6 dB SNR presentation, the recognition score increases from 73% at 0 dB CVR modification to 87% at 6 dB CVR modification and 89% at 12 dB CVR modification.

For the CV6 test, there are no appreciable consonant confusions in the no masking noise case for all the five CVR modifications. For the 12 dB SNR presentation, the consonant confusions are *da/ga* (27%) and *ga/da* (24%) at 0 dB CVR modification; *da/ga* (29%) at 3 dB CVR modification; *da/ga* (33%) and *ga/da* (24%) at 6 dB CVR modification with no appreciable confusions at higher CVRs. For the 6 dB SNR presentation, the consonant confusions are *ka/ta* (25%) and *ka/pa* (24%) at 0 dB CVR modification; *ga/da* (32%) and *ta/ka* (22%) at 3 dB CVR modification; *ga/da* (22%) at 6 dB CVR modification; and *ga/da* (22%) at 9 dB CVR modification. It is observed from the above that there are no voicing errors in the confusions. Furthermore, the number of appreciable consonant confusions is seen to generally decrease with increase in CVR. Examination of the confusion matrices does not reveal any symmetry in these confusions.

The consonant recognition scores obtained by the five subjects for the VC6 test are shown in Table 6.12. Here too, the trend in scores under different conditions for individual subjects was generally similar. Hence, these scores have been averaged across the five subjects as given in the last row of Table 6.12 and plotted in Fig. 6.7. The recognition scores for all the CVR modifications in the no masking noise case, as seen in Table 6.12, are near-perfect. For the 12 dB SNR presentation, the recognition score increases from 85% at 0 dB CVR modification to 97% at 12 dB CVR modification — a total increase of 12%. For the 6 dB SNR presentation, the recognition score increases from 71% at 0 dB CVR modification to 94% at 12 dB CVR modification — an increase of 23%.

For the VC6 test, there are no consonant confusions exceeding 20% in the no masking noise case for all the five CVR modifications. For the 12 dB SNR presentation, the consonant confusions are *ap/at* (23%) at 0 dB CVR modification with no appreciable confusions at higher CVR modifications. For the 6 dB SNR presentation, the consonant confusions are *ap/at* (30%) and *at/ap* (23%) at 0 dB CVR modification; and *ap/at* (23%) at 3 dB CVR modification with no appreciable confusions at higher CVR modifications. It is observed from the above that there are no voicing errors in the confusions considered as appreciable. It is also observed that the number as well as the severity of consonant confusions generally decrease with increasing CVR in both CV and VC contexts, with more positive results in the VC context.

Figs. 6.6 and 6.7 give the plots for the percentage correct recognition scores obtained using the CV6 and VC6 stimuli. The results of the CV6 and VC6 tests thus indicate that increasing CVR does improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects. It is also observed that the recognition scores for stops in the VC context are generally higher than for those in the CV context.

Fig. 6.8 gives the pattern of confusions observed in the identification of place of articulation for different conditions in the CV6 and VC6 tests. For the no masking noise case in the CV6 test, there is no appreciable change in either the pattern of confusions or the overall level of confusions as the CVR is increased from 0 dB to 12 dB. For the 12 dB SNR case too, the pattern of confusions is quite the same at all the CVRs. However, the overall level of confusions is seen to decrease gradually with increasing CVR. For the 6 dB SNR case, the overall level of confusions is found to decrease noticeably for 6 dB CVR modification and thereafter remain unaffected at higher CVRs. The pattern of confusions does not change appreciably with CVR modification.

For the no masking noise case in the VC6 test, near-cent percent recognition scores were obtained for all the CVR modifications. For the 12 dB SNR case, the pattern of confusions appears to remain unaffected with CVR modification. However, the overall level of confusions decreases appreciably and reaches no masking noise levels for CVR modification ≥ 9 dB. In particular, the labial/alveolar (LA/AL) confusions are found to decrease appreciably with increase in CVR. For the 6 dB SNR case, the overall level of confusions are found to decrease gradually with increase in CVR. The pattern of confusions does not change appreciably with increase in CVR.

The consonant recognition scores obtained for the CV6 and VC6 tests were subjected to the paired-t test for statistical significance and the results are tabulated in Table 6.13. Scores under various modification levels were compared to those with no modification. It is observed that increase in C/V ratio is accompanied by a slight improvement in scores ($p<0.4$) in the no masking noise case for the CV6 test. However, for the 12 dB SNR case, the scores improve significantly ($p<0.025$) for CVR modifications of 9 dB and higher. For the 6 dB SNR case, the level of significance in the improvement of- scores is even higher (p<0.001 for 9 dB CVR modification).

Table 6.13 also includes the paired t-test results for the VC6 test. In contrast to the results for the CV6 test, the score is seen to improve significantly (p<0.05) even for a 3 dB CVR modification in the no masking noise case. For the 12 dB SNR case, higher levels of improvement are noted (p<0.005) for 6 dB and higher CVR modifications. For the 6 dB SNR case, significantly much higher levels of improvement (p<0.001) are noted for 9 dB and higher CVR modifications.

Thus the results of the statistical analysis for the CV6 and VC6 tests show that increasing CVR does improve the recognition scores in the presence of masking broadband noise for normal-hearing subjects. It also bears out the earlier observation that the recognition scores for stops in the VC context are generally higher than for those in the CV context.

6.5.2 Information Transmission Analysis

Information transmission analysis results for the CV6 test stimuli, summarized in Table 6.14, show that the overall information transmission as well as the reception of place and voicing features of the consonant in the no masking noise case decreases slightly upto 6 dB CVR modification and then increases slightly for higher CVR modifications. For the 12 dB SNR presentation, the overall information transmission increases from 70% at 0 dB CVR modification to 89% at 12 dB CVR modification. A similar trend is observed in the transmission of the place feature which increases from 50% at 0 dB CVR modification to 81% at 12 dB

CVR modification. There is near-perfect transmission of voicing in both the no masking noise and 12 dB SNR presentations for all the CVR modifications. For the 6 dB SNR presentation, the overall information transmitted increases from 56% at 0 dB CVR modification to 89% at 12 dB CVR modification. The transmission of place information increases from 37% at 0 dB CVR modification to 81% at 12 dB CVR modification while that for voicing increases from 68% at 0 dB CVR modification to 98% at 12 dB CVR modification.

The results of the information transmission analysis are also plotted in Fig. 6.9. For both SNRs of 12 and 6 dB, the transmission of place feature is found to increase appreciably with increasing CVR. It is noted that the information transmission with respect to voicing is higher than that for place feature for all the CVR modifications and SNRs considered.

Information transmission analysis results for the VC6 test stimuli, summarized in Table 6.15, show that the overall information transmission as well as the transmission of voicing feature in the no masking noise case increases with increasing CVR. The transmission of the place feature is near-perfect for the no masking noise case. For the 12 dB SNR presentation, the overall information transmission increases from 69% at 0 dB CVR modification to 92% at 12 dB CVR modification. The information transmitted with respect to place feature increases from 62% at 0 dB CVR modification to 92% and that of voicing feature increases from 69% at 0 dB CVR modification to 89% at 12 dB CVR modification. For the 6 dB SNR presentation, the overall information transmitted increases from 50% at 0 dB CVR modification to 85% at 12 dB CVR modification. The transmission of place and voicing features too follow a similar trend. The transmission of place information increases from 33% at 0 dB CVR modification to 80% at 12 dB CVR modification while that for voicing increases from 67% at 0 dB CVR modification to 82% at 12 dB CVR modification.

The results of the information transmission analysis for the VC6 test are also plotted in Fig. 6.10. For SNRs of 12 and 6 dB, the transmission of place as well as voicing information are found to increase appreciably with CVR modification. A comparison of Figs. 6.9 & 6.10 shows that for all the SNRs considered, while the transmission of information regarding place is superior in the VC context, the transmission of voicing information is superior in the CV context.

6.5.3 Response Time

The response times for the CV6 and VC6 tests, averaged across the five subjects, are summarized in Table 6.16.

For the no masking noise case in the CV6 test, there appears to be an improvement in the response times which more-or-less stabilize for CVR modification ≥ 3 dB. For the two noise situations, there appears to be an improvement in subject response with increasing CVR. For the no masking noise case in the VC6 test, there appears to be an improvent in response times which stabilize for CVR modification ≥ 3 dB. However, the magnitudes of the response times are higher than that for the corresponding situation in the CV6 test. For the two noise cases, however, there appears to be an improvement in subject response times with increasing CVR. Here too, the magnitudes of the response times are higher than in the CV6 test.

The response times, averaged across the five subjects for the CV6 and VC6 tests, for various levels of modification were compared with those for no modification by subjecting the data to the paired t-test for statistical significance. The results are tabulated in Table 6.17. It is observed that, in general, there is a slight improvement ($p<0.4$) in response time with increasing CVR for all the conditions in the two tests.

6.6 CONCLUSIONS

From the results of the CV9 and VC9 tests, it has been observed that increasing CVR does improve recognition scores in the presence of masking broadband noise for normal-hearing subjects. However, the recognition scores for stops in the syllable-final (VC) context are seen to be higher than those for stops in the syllable-initial (CV) position for all the CVR modifications. Similar results have been obtained for the CV6 and VC6 tests too.

The pattern of confusions in the identification of place of articulation remains generally unaffected for all the CVR modifications in all the four tests. Since increase in CVR has been seen to be more effective in bringing down the overall level of confusions in the VC context as compared to that in the CV context, it is suggested that increasing CVR suppresses forward masking of the consonant by the vowel more effectively than backward masking.

Increasing CVR can sometimes result in vowel confusions as seen from the /ip/-/up/ confusions which increased in severity for higher CVR modifications. This confusion was found to increase from 23% at 6 dB CVR modification to 27% at 9 dB CVR modification and 34% at 12 dB CVR modification. Thus, it appears that the possibility of vowel confusions is another factor that could set a limit on the amount of CVR modification that may be used. In experiments with closed-set stimuli, this limit as well as the specific vowel confusions may depend upon the test set involved.

Information transmission analysis results for the CV9 and VC9 tests show that both overall information transmitted as well as transmission of consonant feature increases appreciably with increasing CVR for all the SNRs. However, the overall information transmitted as well as transmission of place feature is seen to be superior in the VC context as compared to those in the CV context. The information transmission of vowel feature for the two tests is seen to be near-perfect for the no masking noise and 12 dB SNR conditions with a slight decrease for the 6 dB SNR condition. Information transmission analysis results for the CV6 and VC6 tests show near-perfect transmission of overall information as well as information transmitted with respect to place and voicing features for the no masking noise case. For the 12 dB and 6 dB SNR cases, the information transmitted with respect to all the three features is found to increase appreciably with increase in CVR modification.

The results of the information transmission analysis for CV6 and VC6 tests further reveal that the information transmitted about place feature is superior in the VC context over that in the CV context. However, the transmission of voicing information is superior in the CV context over that in the VC context.

The average response times for all the four tests do not appear to show any deterioration and in most cases appear to indicate a quicker response by the subjects for increasing values of CVR modification. This has also been borne out by statistical analysis of the results.

These set of experiments were done to study the effect of increasing CVR on stop consonants from three different view-points: recognition score, amount of information transmitted on the basis of feature classification, and the average response time. As seen from the above study, increasing CVR had a positive impact in all the cases. A CVR modification of upto about 10 dB may be used without any adverse effect on vowel recognition for the test stimuli considered here.

TABLE 6.1 Duration of consonant segments and C/V ratios for the synthesized stops /p, t, k, b, d, g/ in the CV and VC contexts of different vowel environments.

Context	Stimulus	Consonant duration, ms	C/V intensity ratio, dB
CV	/pa/	48	-10.2
	/ta/	54	-23.2
	/ka/	49	-17.1
	/pi/	24	-22.1
	/ti/	55	-16.9
	/ki/	39	-25.4
	/pu/	17	-13.5
	/tu/	72	-22.8
	/ku/	48	-19.3
	/ba/	44	- 9.5
	/da/	40	-11.2
	/ga/	43	-12.8
VC	/ap/	29	- 8.1
	/at/	60	-24.1
	/ak/	81	-21.3
	/ip/	26	-16.0
	/it/	54	-10.6
	/ik/	80	-17.4
	/up/	28	- 5.7
	/ut/	54	-16.2
	/uk/	50	-17.4
	/ab/	46	-11.5
	/ad/	42	- 7.0
	/ag/	37	-10.6

TABLE 6.2 Test CV9: Consonant recognition scores, for the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k/ in the CV context of the vowels /a, i, u/. Scores have been averaged across the three vowels. Last row gives the scores and the standard deviations, averaged across the five subjects.

Subject code	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
S1	79	96	83	84	95	47	79	76	79	80	44	52	56	63	83
S2	95	95	90	91	98	56	66	82	80	90	42	59	55	63	73
S3	84	84	84	79	82	52	58	80	82	79	57	60	60	67	74
S4	98	98	100	98	100	81	81	90	84	97	56	65	77	79	83
S5	100	100	100	100	100	69	80	83	87	93	56	59	71	81	87
Avg	**92**	**95**	**92**	**91**	**95**	**62**	**73**	**83**	**82**	**88**	**52**	**59**	**65**	**71**	**80**
SD	9.1	6.4	8.2	9.2	7.4	14.	10.3	5.1	3.3	8.1	7.2	4.7	9.9	8.6	6.2

TABLE 6.3 Test CV9: Consonant recognition scores, averaged across the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k/ in the CV context of the vowels /a, i, u/.

Vowel context	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
/a/	99	100	98	98	100	66	87	87	94	91	60	70	70	88	92
/i/	91	94	90	81	92	68	77	87	84	91	55	69	75	77	79
/u/	85	90	86	93	92	51	55	73	69	82	39	40	48	48	69

TABLE 6.4 Test VC9: Consonant recognition scores, for the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k/ in the VC context of the vowels /ɑ, i, u/. Scores have been averaged across the three vowels. Last row gives the scores and the standard deviations, averaged across the five subjects.

Subject code	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
S1	100	100	100	100	100	84	95	98	96	96	75	83	87	84	86
S2	97	100	100	99	98	76	95	93	90	99	76	76	84	85	95
S3	100	100	100	100	100	87	93	96	98	99	59	72	82	92	99
S4	100	100	100	99	100	84	83	84	100	97	76	86	79	91	88
S5	100	100	100	100	100	94	99	98	98	100	80	85	96	96	98
Avg	**99**	**100**	**100**	**100**	**100**	**85**	**93**	**93**	**97**	**98**	**73**	**81**	**85**	**89**	**93**
SD	1.3	0	0	0.4	1.0	6.4	5.9	5.4	4.0	1.6	8.0	6.3	6.6	4.9	6.0

TABLE 6.5 Test VC9: Consonant recognition scores, averaged across the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k/ in the VC context of the vowels /ɑ, i, u/.

Vowel context	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
/ɑ/	100	100	100	100	100	85	91	95	98	96	78	86	92	98	99
/i/	100	100	100	99	100	90	95	91	96	98	76	81	81	87	86
/u/	98	100	100	100	99	79	90	92	94	97	65	74	82	84	92

TABLE 6.6 Tests CV9 and VC9: Paired t-test significance levels at different C/V ratio modifications (CVRMs) with reference to 0 dB CVRM for the consonant recognition scores of Tables 6.2 and 6.4 averaged across the five subjects. N=5.(NS = not significant).

Test	SNR (dB)	CVRM (dB)	mean	SD	Test of difference (two-tailed)	
					t	p
CV9	No masking noise	0	92	9.1		
		3	95	6.4	0.68	NS
		6	92	8.2	0.01	NS
		9	91	9.2	0.14	NS
		12	95	7.4	0.71	<0.5
	12	0	62	14.0		
		3	73	10.3	1.06	<0.4
		6	83	5.1	3.21	<0.025
		9	82	3.3	3.37	<0.01
		12	88	8.1	3.75	<0.01
	6	0	52	7.2		
		3	59	4.7	2.06	<0.1
		6	65	9.9	2.35	<0.05
		9	71	8.6	3.83	<0.005
		12	80	6.2	6.76	<0.001
VC9	No masking noise	0	99	1.3		
		3	100	0.0	1.00	<0.4
		6	100	0.0	1.00	<0.4
		9	100	0.4	0.31	NS
		12	100	1.0	0.28	NS
	12	0	85	6.4		
		3	93	5.9	2.03	<0.1
		6	93	5.4	2.25	<0.1
		9	97	4.0	3.38	<0.01
		12	98	1.6	4.41	<0.005
	6	0	73	8.0		
		3	81	6.3	1.58	<0.2
		6	85	6.6	2.66	<0.05
		9	89	4.9	3.82	<0.01
		12	93	6.0	4.44	<0.005

TABLE 6.7 Test CV9: Information transmission analysis for the stops /p, t, k/ in the CV context of the three vowels /a, i, u/.

Feature	Relative information transmitted (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
Overall	92 88	92	89	88	92	59	70	77	77	82	49	57	56	63	73
Con-sonant	72	79	71	69	81	17	32	48	48	62	8	17	23	32	48
Vowel	99	100	100	100	98	94	97	98	99	97	83	85	80	80	88

TABLE 6.8 Test VC9: Information transmission analysis for the stops /p, t, k/ in the VC context of the three vowels /a, i, u/.

Feature	Relative information transmitted (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
Overall	99	100	100	99	99	77	87	88	93	97	66	71	77	83	90
Con-sonant	97	100	100	98	98	55	75	81	88	97	40	52	67	78	96
Vowel	99	100	100	99	100	95	96	91	96	94	88	84	83	81	81

TABLE 6.9 Response times in sec (and standard deviations) averaged across the five subjects for the CV9 and VC9 tests.

Test	SNR (dB)	Response time (sec)				
		CVRM (dB)				
		0	3	6	9	12
CV9	No masking noise	2.09 (0.43)	2.12 (0.57)	2.05 (0.66)	2.12 (0.60)	2.09 (0.61)
	12	2.57 (0.69)	2.79 (1.16)	2.34 (0.79)	2.21 (0.51)	2.30 (0.68)
	6	2.61 (0.89)	2.50 (0.64)	2.47 (0.55)	2.50 (0.66)	2.36 (0.75)
VC9	No masking noise	2.01 (0.40)	1.87 (0.37)	1.84 (0.44)	2.38 (1.33)	1.90 (0.40)
	12	2.29 (0.69)	2.11 (0.44)	2.10 (0.49)	2.09 (0.59)	2.23 (0.63)
	6	2.35 (0.55)	2.33 (0.67)	2.31 (0.60)	2.35 (0.62)	2.02 (0.42)

TABLE 6.10 Tests CV9 and VC9: Paired t-test significance levels at different C/V ratio modifications (CVRMs) with reference to 0 dB CVRM for the response times (in sec) of Table 6.9 averaged across the five subjects. N=5. (NS = not significant).

Test	SNR (dB)	CVRM (dB)	mean	SD	Test of difference (two-tailed)	
					t	p
CV9	No masking noise	0	2.09	0.43		
		3	2.12	0.57	0.09	NS
		6	2.05	0.66	0.11	NS
		9	2.12	0.60	0.09	NS
		12	2.09	0.61	0.00	NS
	12	0	2.57	0.69		
		3	2.79	1.16	0.40	NS
		6	2.34	0.79	0.49	NS
		9	2.21	0.51	0.94	<0.4
		12	2.30	0.68	0.62	NS
	6	0	2.61	0.89		
		3	2.50	0.64	0.22	NS
		6	2.47	0.55	0.30	NS
		9	2.50	0.66	0.22	NS
		12	2.36	0.75	0.48	NS
VC9	No masking noise	0	2.01	0.40		
		3	1.87	0.37	0.57	NS
		6	1.84	0.44	0.64	NS
		9	2.38	1.33	0.60	NS
		12	1.90	0.40	0.43	NS
	12	0	2.29	0.89		
		3	2.11	0.44	0.49	NS
		6	2.10	0.49	0.50	NS
		9	2.09	0.59	0.49	NS
		12	2.23	0.63	0.14	NS
	6	0	2.35	0.55		
		3	2.33	0.67	0.05	NS
		6	2.31	0.60	0.11	NS
		9	2.35	0.62	0.00	NS
		12	2.02	0.42	2.28	<0.1

TABLE 6.11 Test CV6: Consonant recognition scores, for the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/. Last row gives the scores and the standard deviations, averaged across the five subjects.

Subject code	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
S1	100	100	100	100	100	69	97	89	91	96	64	94	80	90	96
S2	97	91	97	99	100	80	89	91	96	99	82	82	90	94	89
S3	94	91	91	87	87	74	89	83	90	86	70	70	83	93	83
S4	100	100	100	100	100	86	92	91	97	92	78	67	90	86	90
S5	100	91	91	87	92	91	83	89	89	97	72	88	91	91	91
Avg	**98**	**95**	**94**	**95**	**96**	**81**	**90**	**87**	**91**	**94**	**73**	**83**	**87**	**90**	**89**
SD	2.6	4.9	4.2	7.1	6.1	9.0	4.8	4.	2 6	5.2	6.5	9.0	4.7	5.0	4.

TABLE 6.12 Test VC6: Consonant recognition scores, for the five subjects, under different C/V ratio modifications (CVRMs) and SNRs, for the stops /p, t, k, b, d, g/ in the VC context of the vowel a . Last row gives the scores and the standard deviations, averaged across the five subjects.

Subject code	Consonant recognition scores (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
S1	89	99	99	100	100	82	86	96	99	93	66	77	89	92	94
S2	93	97	98	97	99	86	91	90	97	100	70	83	86	97	89
S3	87	100	100	100	100	77	74	96	89	93	63	67	80	87	91
S4	96	99	100	100	100	89	91	98	97	100	76	84	90	96	97
S5	99	100	100	100	100	89	90	98	98	99	80	82	90	96	98
Avg	**93**	**99**	**99**	**99**	**100**	**85**	**86**	**95**	**96**	**97**	**71**	**79**	**87**	**93**	**94**
SD	4.9	1.3	1.0	1.5	0.5	5.2	7.1	3.	4.0	3.5	6.9	7.3	4.3	4.1	3.

TABLE 6.13 Tests CV6 and VC6: Paired t-test significance levels at different C/V ratio modifications (CVRMs) with reference to 0 dB CVRM for the consonant recognition scores of Tables 6.11 and 6.12 averaged across the five subjects. N=5. (NS = not significant).

Test	SNR (dB)	CVRM (dB)	mean	SD	Test of difference (two-tailed)	
					t	p
CV6	No masking noise	0	98	2.6		
		3	95	4.9	1.43	<0.2
		6	94	4.2	1.02	<0.4
		9	95	7.1	1.08	<0.4
		12	96	6.1	0.96	<0.4
	12	0	81	9.0		
		3	90	4.8	2.19	<0.1
		6	87	4.0	2.03	<0.1
		9	91	2.6	2.94	<0.025
		12	94	5.2	3.06	<0.025
	6	0	73	6.5		
		3	83	9.0	2.15	<0.1
		6	87	4.7	3.54	<0.01
		9	90	5.0	5.13	<0.001
		12	89	4.8	4.41	<0.005
VC6	No masking noise	0	93	4.9		
		3	99	1.3	2.73	<0.05
		6	99	1.0	2.95	<0.025
		9	99	1.5	2.89	<0.025
		12	100	0.5	3.17	<0.025
	12	0	85	5.2		
		3	86	7.1	0.45	NS
		6	95	3.2	4.04	<0.005
		9	96	4.0	3.99	<0.005
		12	97	3.5	4.39	<0.005
	6	0	71	6.9		
		3	79	7.3	1.71	<0.2
		6	87	4.3	4.37	<0.005
		9	93	4.1	6.21	<0.001
		12	94	3.7	6.37	<0.001

TABLE 6.14 Test CV6: Information transmission analysis for the stops /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/.

Feature	Relative information transmitted (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
Overall	96	91	87	91	93	70	84	80	84	89	56	75	75	82	89
Place	93	84	81	82	86	50	74	67	78	81	37	63	63	73	81
Voicing	100	98	92	95	100	97	95	88	87	98	68	89	82	87	98

TABLE 6.15 Test VC6: Information transmission analysis for the stops /p, t, k, b, d, g/ in the VC context of the vowel /ɑ/.

Feature	Relative information transmitted (%)														
	No masking noise					SNR = 12 dB					SNR = 6 dB				
	CVRM (dB)					CVRM (dB)					CVRM (dB)				
	0	3	6	9	12	0	3	6	9	12	0	3	6	9	12
Overall	86	97	98	98	99	69	71	89	90	92	50	61	70	84	85
Place	94	98	99	97	99	62	64	86	93	92	33	47	62	82	80
Voicing	67	94	96	100	100	69	67	84	79	89	67	65	70	80	82

TABLE 6.16 Response times in sec (and standard deviations) averaged across the five subjects for the CV6 and VC6 tests.

Test	SNR (dB)	Response time (sec)				
		CVRM (dB)				
		0	3	6	9	12
CV6	No masking noise	2.04 (0.65)	1.63 (0.72)	1.67 (0.65)	2.11 (1.30)	1.65 (0.58)
	12	2.08 (0.62)	2.01 (0.56)	2.18 (0.71)	1.89 (0.45)	1.82 (0.54)
	6	2.29 (0.56)	2.11 (0.73)	2.03 (0.50)	1.95 (0.61)	2.04 (0.64)
VC6	No masking noise	2.59 (0.35)	2.24 (0.44)	2.22 (0.51)	2.17 (0.52)	2.18 (0.35)
	12	2.58 (0.37)	2.50 (0.37)	2.23 (0.38)	2.27 (0.35)	2.19 (0.23)
	6	2.74 (0.35)	2.80 (0.41)	2.58 (0.29)	2.40 (0.46)	2.36 (0.44)

TABLE 6.17 Tests CV6 and VC6: Paired t-test significance levels at different C/V ratio modifications (CVRMs) with reference to 0 dB CVRM for the response times (in sec) of Table 6.16 averaged across the five subjects. N=5. (NS = not significant).

Test	SNR (dB)	CVRM (dB)	mean	SD	Test of difference (two-tailed)	
					t	p
CV6	No masking noise	0	2.04	0.65		
		3	1.63	0.72	0.95	<0.4
		6	1.67	0.65	0.90	<0.4
		9	2.11	1.30	0.11	NS
		12	1.65	0.58	1.00	<0.4
	12	0	2.08	0.62		
		3	2.01	0.56	0.19	NS
		6	2.18	0.71	0.24	NS
		9	1.89	0.45	0.55	NS
		12	1.82	0.54	0.71	<0.5
	6	0	2.29	0.56		
		3	2.11	0.73	0.44	NS
		6	2.03	0.50	0.78	<0.5
		9	1.95	0.61	0.92	<0.4
		12	2.04	0.64	0.66	NS
VC6	No masking noise	0	2.59	0.35		
		3	2.24	0.44	0.99	<0.4
		6	2.22	0.51	1.34	<0.4
		9	2.17	0.52	1.50	<0.2
		12	2.18	0.35	1.85	<0.2
	12	0	2.58	0.37		
		3	2.50	0.37	0.34	NS
		6	2.23	0.38	1.48	<0.2
		9	2.27	0.35	1.36	<0.4
		12	2.19	0.23	2.00	<0.1
	6	0	2.74	0.35		
		3	2.80	0.41	0.25	NS
		6	2.58	0.29	0.79	<0.5
		9	2.40	0.46	1.32	<0.4
		12	2.36	0.44	1.51	<0.2

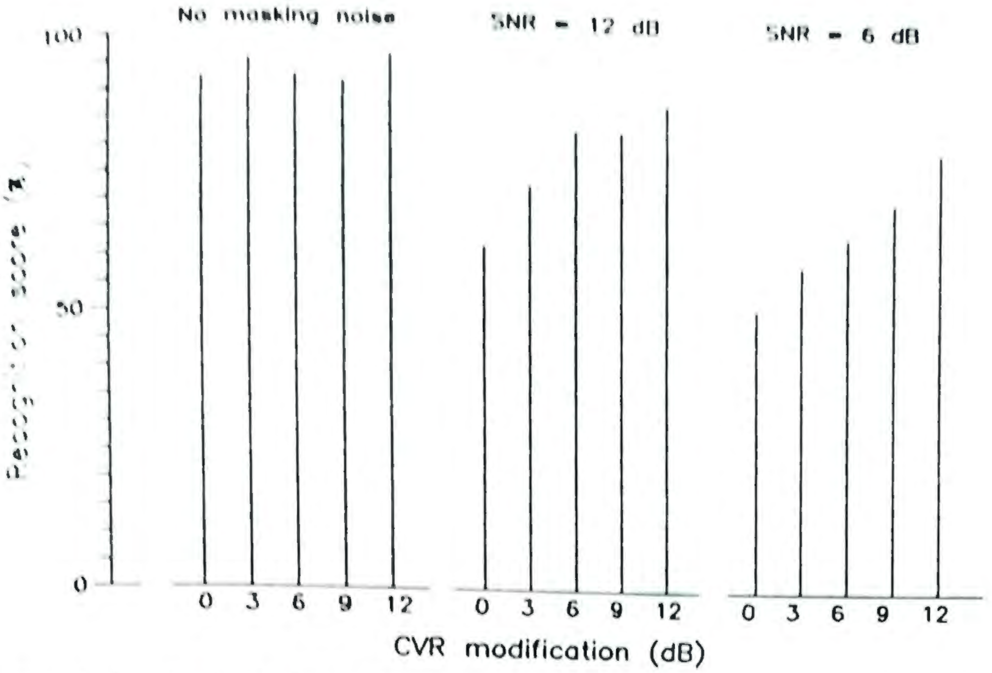


FIG. 6.1 Test CV9: Percentage recognition scores for the stops /p, t, k/ in the CV context averaged across the three vowels /a, i, u/.

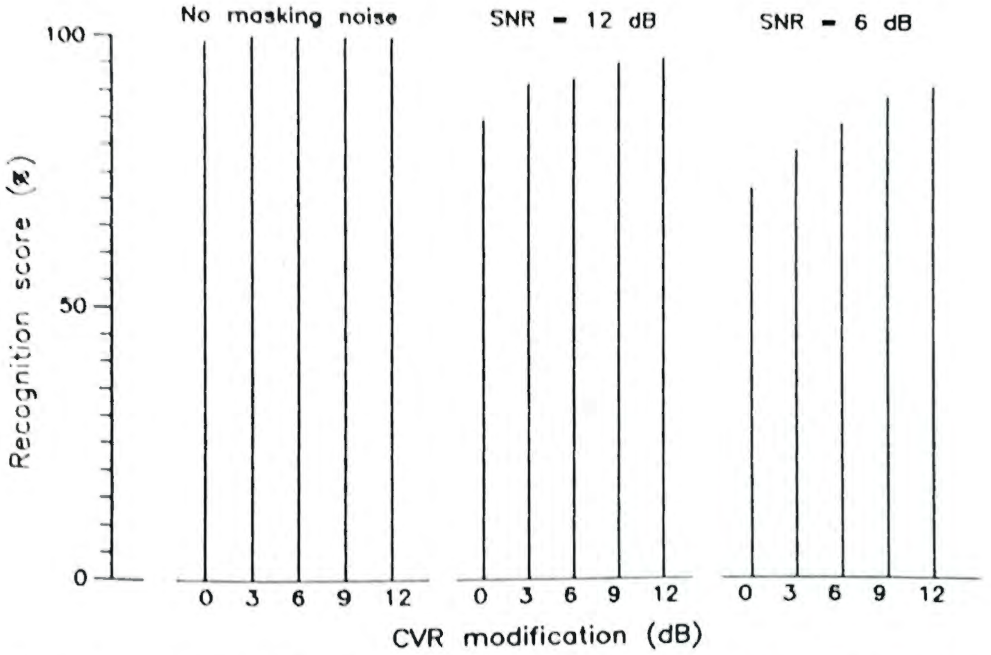


FIG. 6.2 Test VC9: Percentage recognition scores for the stops /p, t, k/ in the VC context averaged across the three vowels /a, i, u/.

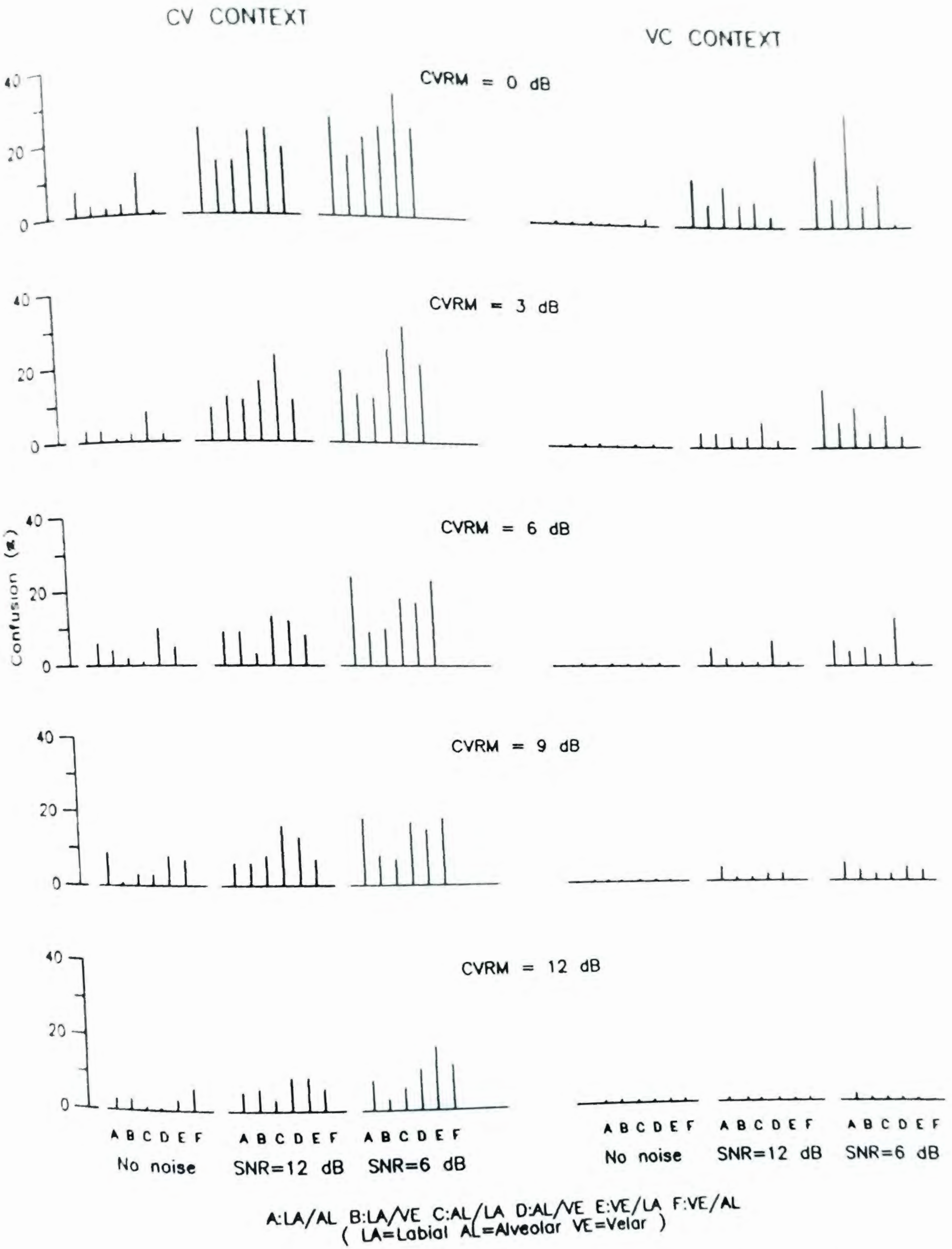


FIG. 6.3 Place of articulation confusions for the stops /p, t, k/ in the CV and VC contexts of the vowels /a, i, u/.

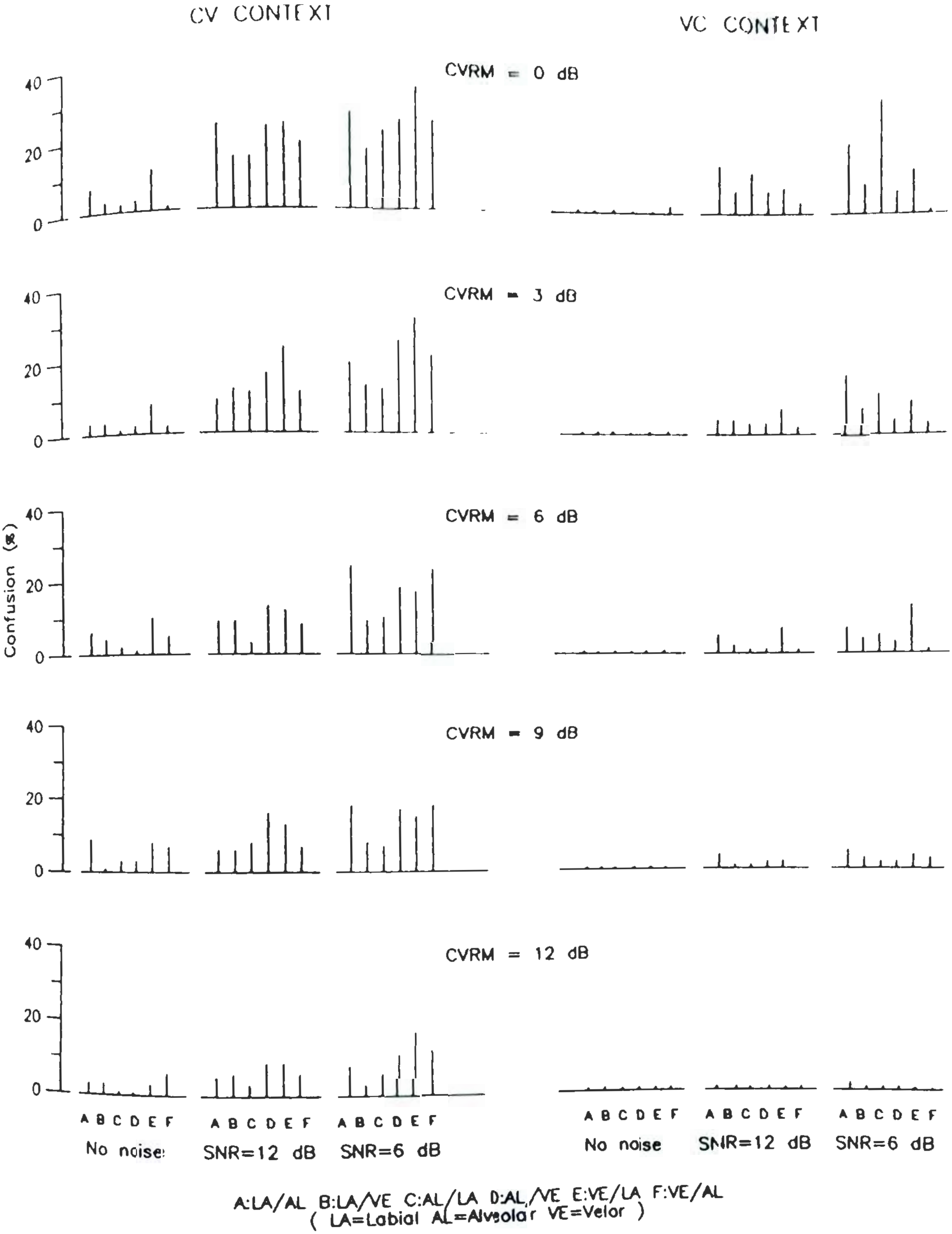
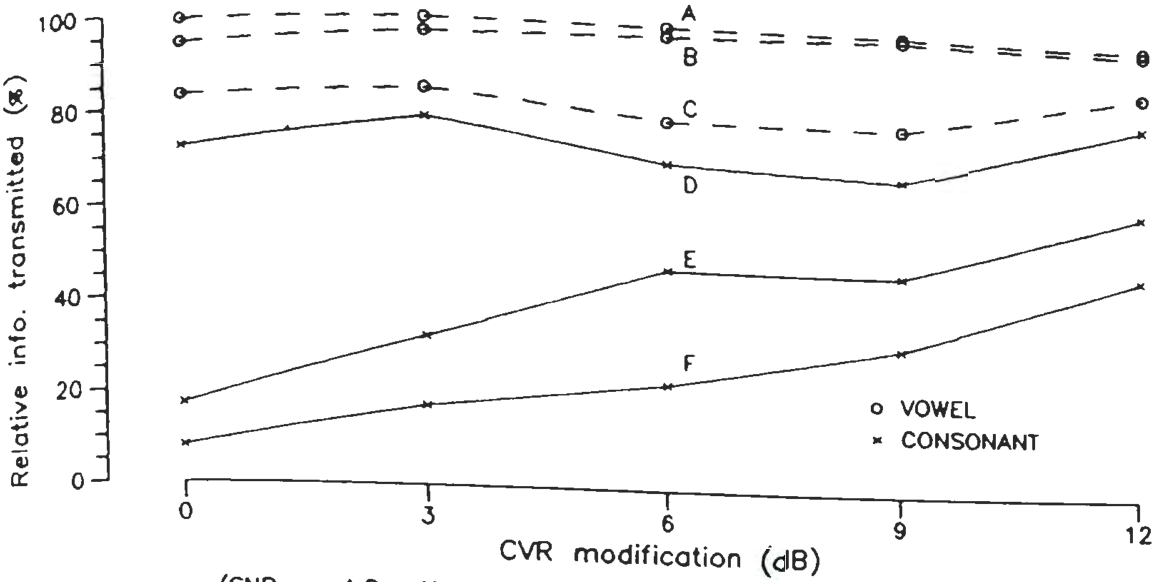
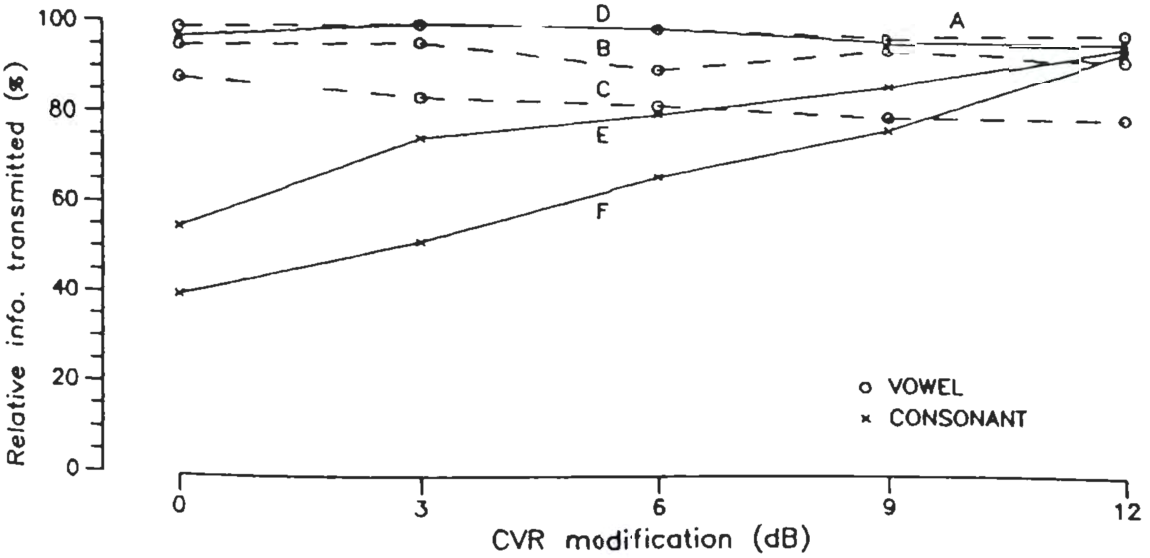


FIG. 6.3 Place of articulation confusions for the stops /p, t, k/ in the CV and VC contexts of the vowels /a, i, u/.



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 6.4 Test CV9: The relative information transmitted about consonant and vowel features plotted as a function of CVR modification for different SNRs in CV context.



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 6.5 Test VC9: The relative information transmitted about consonant and vowel features plotted as a function of CVR modification for different SNRs in VC context.

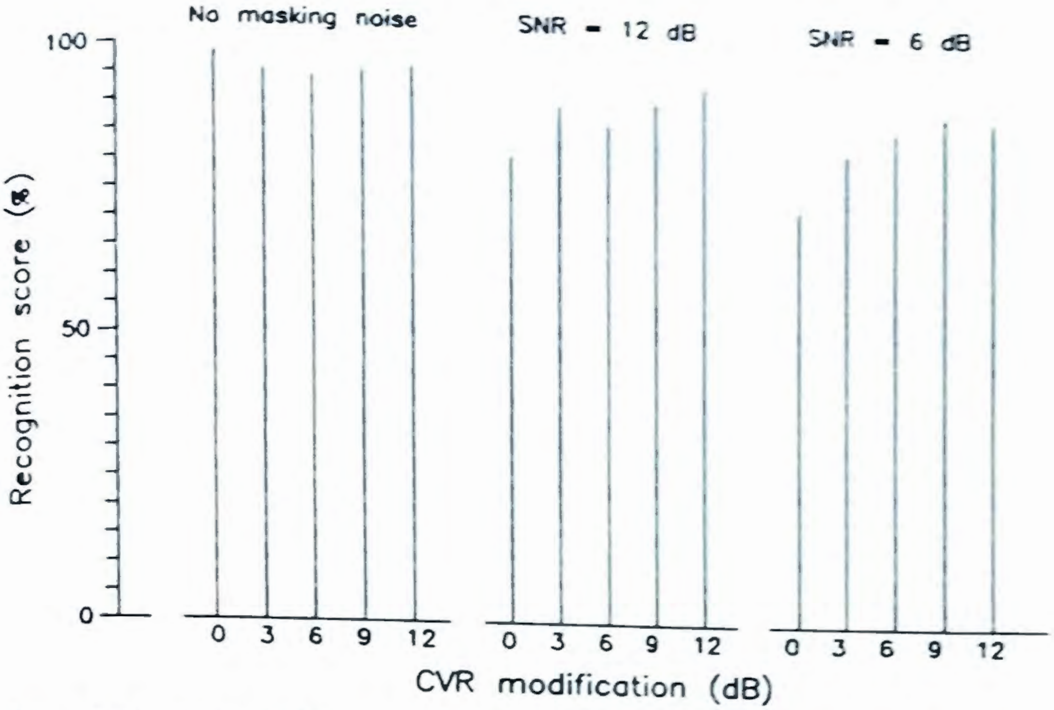


FIG. 6.6 Test CV6: Percentage recognition scores for the stops /p, t, k, b, d, g/ in the CV context of the vowel /a/.

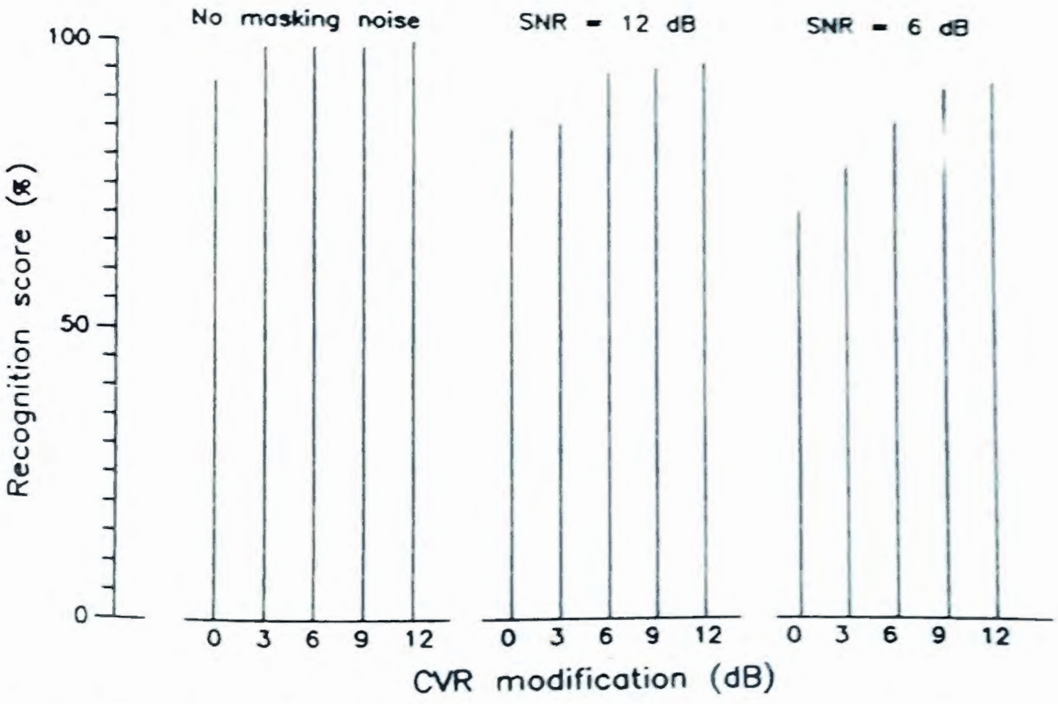


FIG. 6.7 Test VC6: Percentage recognition scores for the stops /p, t, k, b, d, g/ in the VC context of the vowel /a/.

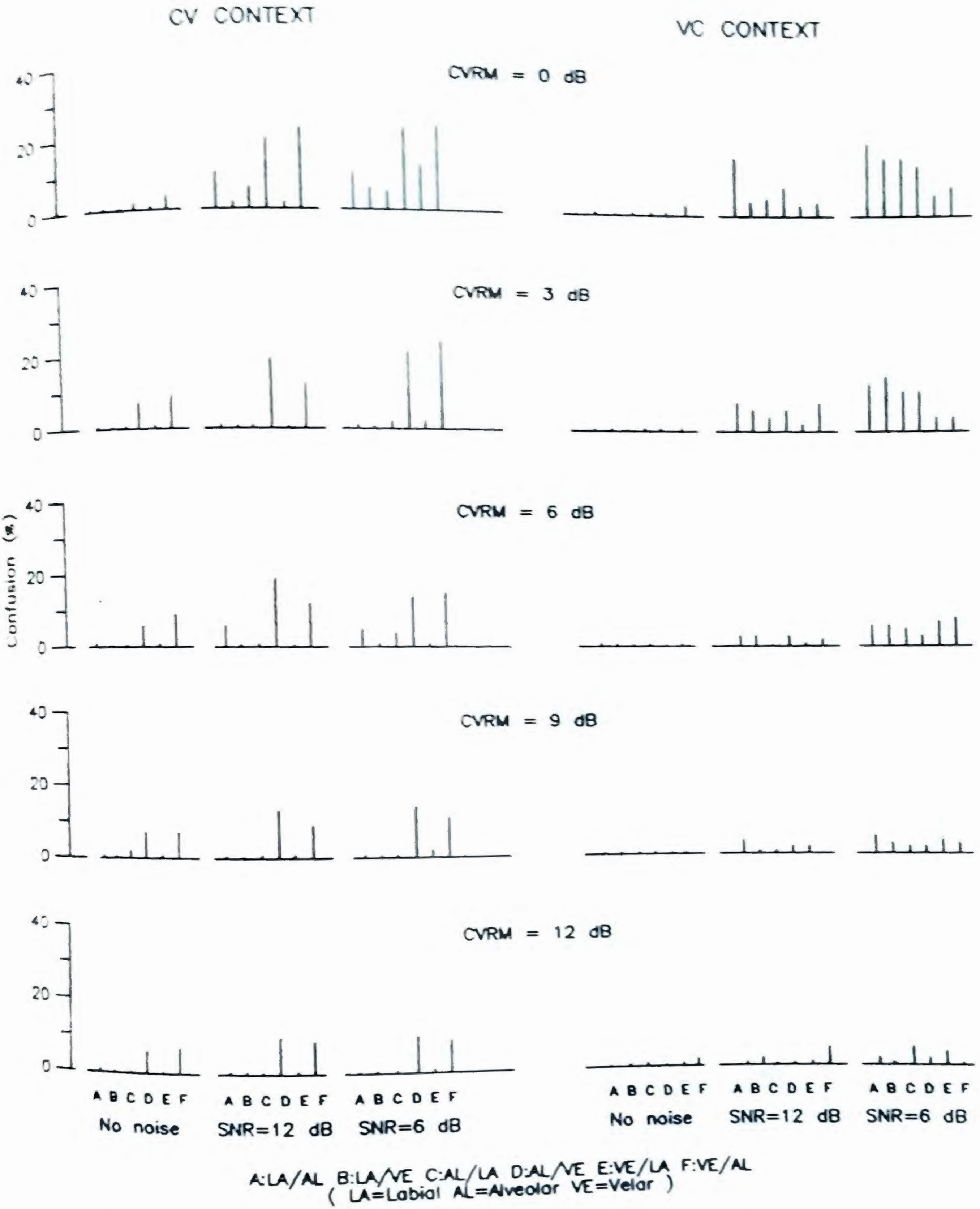
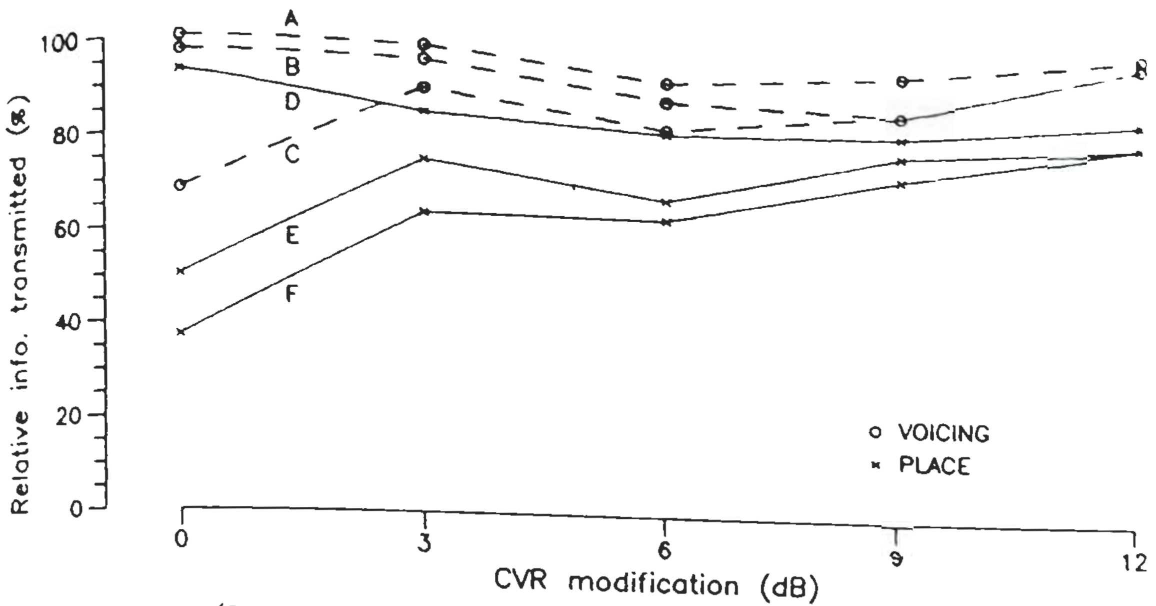
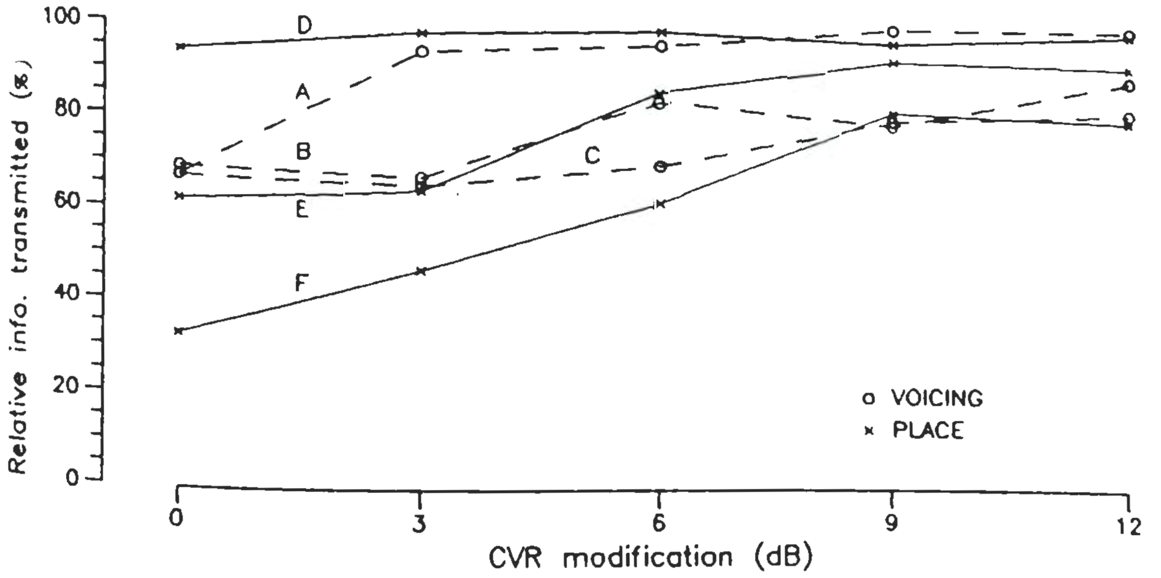


FIG. 6.8 Place of articulation confusions for the stops /p, t, k, b, d, g/ in the CV and VC contexts of the vowel /a/.



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 6.9 Test CV6: The relative information transmitted about place and voicing features plotted as a function of CVR modification for different SNRs in CV context.



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 6.10 Test VC6: The relative information transmitted about place and voicing features plotted as a function of CVR modification for different SNRs in VC context.

CHAPTER 7

EFFECT OF CONSONANT DURATION MODIFICATION

7.1 INTRODUCTION

The previous chapter had presented the results of experiments performed to study the effect of consonant-to-vowel intensity ratio (CVR) modification on the recognition of stop consonants. In the present chapter, the results and implications of experiments performed on stimuli with increased consonant duration are presented.

7.2 TEST STIMULI

The aim of the experiments was to study the effects of increasing consonant duration on the recognition of stop consonants. As discussed earlier in Section 4.2, the articulation and perception of stop consonants is associated with various acoustic segments including closure, burst, formant transition, and voice onset time (VOT). All these segments contribute to the increased duration in clear speech (Section 3.4). It was decided to synthesize stimuli in which each of the acoustic segments was increased in duration by different amounts keeping the durations of the other acoustic segments unaltered. This was done in order to study the effects of altering the duration of each of these acoustic segments separately.

As in the case of the CVR modification experiments, only isolated syllables with stop consonants in the CV and VC contexts were considered for studying the effect of consonant duration modification. Since only isolated syllables are used as stimuli, the initial silence or closure segment cannot be defined for syllables in the CV context. For syllables in the VC context, since there is no vowel following the consonant, VOT cannot be studied as an acoustic segment of perceptual importance. Keeping these points in mind, the six stop consonants /p, t, k, b, d, g/ in the CV context of the vowel /*a*/ have been chosen as the stimuli to study the following:

1) The effect of burst duration (BD) modification on consonant recognition. For this, a new version of each syllable was synthesized by increasing the burst duration by 100%. The C/V intensity ratio (CVR), which would have got affected in the process, was kept within ±1 dB of the normal value by adjusting the vowel amplitude (AV) during synthesis. Thus each syllable had two versions, one with no modification and the other with BD modification of 100%.
2) The effect of modifying formant transition duration (FTD). For this, two new versions of each syllable were synthesized by increasing the formant transition durations by 50% and 100%. To neutralize the effect of this modification on the CVR, the vowel amplitude was adjusted during synthesis. Thus each syllable had three versions, one with no modification and the other two with FTD modification of 50% and 100%.
3) The effect of modifying VOT. Two new versions of each syllable were synthesized by increasing the VOT by 50% and 100%. Here too, CVR was kept constant by adjusting the vowel

amplitude. Thus each syllable had three versions, one with no modification and the other two with VOT modification of 50% and 100%.

The modifications in duration were limited to the above range in order not to lose the perceptual characteristics of the consonants. In all the cases, the overall duration of each CV syllable was maintained at 300 ms by adjusting the vowel duration appropriately. Hearing impairment was simulated in normal-hearing listeners by mixing each stimulus with synthesized broadband noise. As in the CVR modification experiments, the tests were performed under three listening conditions: no masking noise, and masking noise with 12 dB SNR and 6 dB SNR. Thus, the BD modification test included 6 CV syllables under 6 conditions (2 BD modifications X 3 SNRs) with a total of 36 stimuli. The FTD modification test included 6 CV syllables under 9 conditions (3 FTD modifications X 3 SNRs), thus involving 54 stimuli. The VOT modification test too involved 6 CV syllables under 9 conditions (3 VOT modifications X 3 SNRs) yielding 54 stimuli. Appendix B includes sample spectrograms of duration modified stimuli.

7.3 EXPERIMENTAL METHOD

Four of the five subjects who participated in the CVR modification experiments, were available for the duration modification experiments.

The apparatus and the presentation procedure were the same as those adopted for the CVR modification experiments. For each subject, the BD modification test involving 6 experimental conditions took about three to four hours for completion. The FTD and VOT modification tests which involved 18 experimental conditions between them took about eight to ten hours together for completion. Thus the three tests required about twelve to fourteen hours per subject. The experiments were spread over a one-month period for the four subjects tested.

Just as in the CVR modification experiments, the test runs were randomized in order to reduce biases due to learning effects. At the end of each run, the stimulus-response confusion matrix, the recognition score, and the average response time were stored. For evaluation purposes, the stimulus-response confusion matrices from a number of test runs for each experimental condition were combined, and two different measures based on them were used, namely, recognition score and relative information transmission. In addition, average response time was also considered as a possible measure of comparing the test stimuli processed differently.

7.4 RESULTS

The results for each of the three tests are presented here. The recognition scores are considered first, followed by information transmission analysis, and average response times. Although useful and easy to interpret, the recognition score does not provide any information on the distribution of errors. Information transmission analysis has the merit that it measures the covariance between the stimuli and responses and hence takes into account the relatedness of the two. The test stimuli processed differently could also be compared on the basis of average response time, with increasing response time generally indicating an increasing level of difficulty encountered by the subject in processing.

7.4.1 Recognition Scores

The consonant recognition scores obtained for the burst duration (BD) modification test are given in Table 7.1. All the four subjects show a reduction in scores as SNR is reduced for both unmodified stimuli as well as stimuli whose burst duration is doubled. For the case of no masking noise, Subjects S1 and S2 show near-perfect recognition scores which decrease by about 3% when burst duration is doubled. However, the score for Subject S3 shows an increase of 8% when burst duration is doubled while Subject S4 shows perfect identification for both the burst duration modifications.

For the 12 dB SNR presentation, while S1 and S4 show a 2% and 4% decrease in score, S2 and S3 show a 12% and 1% increase in score. For the 6 dB SNR presentation, while S1, S2, and S3 show a 4%, 7%, and 8% increase in score, S4 shcws a 7% decrease in score when burst duration is doubled.

The recognition scores for the BD modification test, averaged across the four subjects, are seen in the last row of Table 7.1 and plotted in Fig. 7.1. In the case of no masking noise, the recognition score is unaffected when burst duration is doubled. For the 12 dB and 6 dB SNR presentations, a very slight increase (1%) in recognition score is observed when burst duration is doubled.

For the BD modification test, there are no appreciable consonant confusions in the no masking noise case. For the 12 dB SNR case, the consonant confusions are *da/ga* (22%) for 0% BD increase with no appreciable confusions for 100% BD increase. For the 6 dB SNR case, the consonant confusions are *ka/ta* (34%), *ta/ka* (23%), and *ga/da* (23%) for 0% BD increase; and *ga/da* (40%) and *ta/ka* (32%) for 100% BD increase.

The consonant recognition scores for the test involving formant transition duration (FTD) modification are given in Table 7.2. All the four subjects exhibit a reduction in scores as SNR is reduced for all the FTD modifications. For the case of no masking noise, subjects S1 and S2 show near-perfect recognition scores with their scores decreasing by 3% and 12% when FTD is doubled, and S4 shows perfect identification. The scores for S3 are slightly lower and are seen to decrease by 2% when FTD is doubled.

For the 12 dB SNR case, while S1, S2, and S4 show a 3%, 15%, and 11% decrease in scores when transition duration is doubled, S3 registers a 13% increase. For 6 dB SNR S1, S2, and S3 show a 16%, 4%, and 16% increase respectively for 50% FTD increase and then a 6%, 9%, and 5% decrease for 100% FTD increase. The score for S4 decreases by 11% as FTD is doubled.

The recognition scores, averaged across the four subjects, for the FTD modification test, are seen in the last row of Table 7.2 and plotted in Fig. 7.2. The score decreases by 3% in the no masking noise case and by 4% in the 12 dB SNR case when FTD is doubled. For the 6 dB SNR case, it first increases by 4% for 50% FTD increase and then falls by 4% for 100% FTD increase.

For the FTD modification test, consonant confusions in the no masking noise case include da/ga (25%) for 50% FTD increase and da/ga (23%) for 100% FTD increase. For the 12 dB SNR case, the consonant confusions are ga/da (26%) and ta/ka (23%) for 0% FTD increase; da/ga (43%) and ga/da (27%) for 50% FTD increase; and da/ga (60%) and ta/da (40%) for 100% FTD increase. For the 6 dB SNR case, the consonant confusions are ka/ta (30%), ta/ka (29%), and ga/da (22%) for 0% FTD increase; da/ga (43%), ga/da (33%), and ta/da (22%) for 50% FTD increase; and da/ga (54%), ta/ka (49%), and ka/ta (23%) for 100% FTD increase.

The consonant recognition scores for the voice onset time (VOT) modification test are given in Table 7.3. All four subjects exhibit a reduction in scores as SNR is decreased for all the VOT modifications. For the no masking noise case, while the scores for S1, S2, and S4 are near-perfect and decrease by 6%, 9%, and 3% when VOT is doubled, the score for S3 is seen to increase by 8% for 50% VOT increase and then fall by 10% for 100% VOT increase. For the 12 dB SNR case, while S1 shows a 6% increase in score as VOT is doubled, S2, S3, and S4 show a 22%, 7%, and 3% decrease respectively. For the 6 dB SNR case, while S1 shows an 8% increase in score as VOT is doubled, S2 and S3 show a 21% and 7% decrease respectively. The score for S4 remains fairly unaffected with increase in VOT.

The recognition scores, averaged across the four subjects, for the VOT modification test, are seen in the last row of

Table 7.3 and plotted in Fig. 7.3. A 4% decrease in score is seen when VOT is doubled in the no masking noise case. For the 12 dB SNR case, the score falls by 6% and for the 6 dB SNR case it falls by 4% as VOT is doubled.

For the VOT modification test, there are no appreciable consonant confusions in the no masking noise case for all the FTD modifications. For the 12 dB SNR case, the consonant confusions include d*a*/g*a* (28%) for 0% VOT increase; and t*a*/k*a* (42%) for 50% VOT increase. For the 6 dB SNR case, the consonant confusions are k*a*/t*a* (43%) and t*a*/k*a* (30%) for 0% VOT increase; t*a*/k*a* (33%) and k*a*/t*a* (28%) for 50% VOT increase; and p*a*/t*a* (23%) and t*a*/k*a* (22%) for 100% VOT increase.

The consonant recognition scores obtained for the above three tests, averaged across the four subjects, were subjected to the paired t-test for statistical significance and tabulated in Table 7.4. Scores under various modification levels were compared to those with no modification. No significant improvement is observed when burst duration is doubled. However, for the FTD modification test, a slight improvement in performance ($p<0.5$) is observed as FTD is increased by 50% for the 6 dB SNR case. When FTD is increased by 100%, no significant change in performance is seen for the no masking noise and 6 dB SNR conditions. However, there is a slight degradation in performance ($p<0.4$) in the 12 dB SNR case.

For the VOT modification test, there is no significant change in performance when VOT is increased by 50% in both the no masking noise as well as the 6 dB SNR cases. There is a slight degradation in performance (p<0.5) when VOT is doubled in the no masking noise case, while in the 6 dB SNR case no significant change in performance is observed. For the 12 dB SNR case, a slightly significant degradation in performance (p<0.2) is noted when VOT is increased by 50%. However, the level of degradation is seen to be slightly lower (p<0.4) when VOT is doubled.

The results of the statistical analysis suggest that of the three acoustic segments contributing to increased consonant duration in clear speech, formant transition duration modification yields slightly positive results.

7.4.2 Information Transmission Analysis

Information transmission analysis results for overall information transmission as well as for consonant place and voicing feature classifications of the BD modification test stimuli are summarized in Table 7.5. For the no masking noise case, S4 shows perfect overall information transmission for both the BD modifications. For S1 and S2, the overall information transmitted is near-perfect and decreases by 4% and 6%, while that for S3 is lower and increases by 2% when burst duration is doubled. Transmission of place information is seen to increase for S1, S2, and S3 by 8%, 4%, and 14% when burst duration is doubled. Howevever, the transmission of voicing information is found to decrease for S1, S2, and S3 by'30%, 26%, and 8% respectively.

For the 12 dB SNR case in the BD modification test, S1, S3, and S4 show a show a decrease in overall information transmitted as well as transmission of place information when burst duration is doubled. However, the voicing information transmission is found to decrease by 18% for S1 and 13% for S3 and to increase by 5% for S4. For S2, the overall information transmitted as well as transmission of place and voicing features increase by 13%, 22%, and 8% respectively. For the 6 dB SNR case, the overall information transmitted increases for S1, S2, and S3 by 1%, 6%, and 9% respectively and decreases for S4 by 6%. The transmission of place information increases for S1, S2, and S3 by 8%, 14%, and 12% respectively and decreases for S4 by 7% when burst duration is doubled. The information according to voicing decreases for S1, S2, and S4 by 13%, 13%, and 11% respectively and increases for S3 by 14%.

The relative information transmitted, averaged across the four subjects. for the burst duration modification test, are seen in the last row for each feature in Table 7.5. The overall information transmitted decreases by 2% in the no masking noise case, remains unchanged in the 12 dB SNR case, and increases by 2% in the 6 dB SNR case. The transmission of place information increases by 7% in the no masking noise case, 2% in the 12 dB SNR case, and 5% in the 6 dB SNR case when burst duration is doubled. However, the transmission of voicing information decreases by 5% in the no masking noise case, 3% in the 12 dB SNR case, and 5% in the 6 dB SNR case when burst duration is doubled. These results are plotted in Fig. 7.4.

Information transmission analysis results for the formant transition duration (FTD) modification test stimuli are summarized in Table 7.6. For the no masking noise case, the overall information transmitted decreases for S1, S2, and S3 by 5%, 12%, and 2% respectively when FTD is doubled while S4 shows near-perfect information transmission. Transmission of place information is seen to decrease for S1, S2, and S3 by 11%, 26%, and 4% respectively and remain perfect for S4 when FTD is doubled. All four subjects exhibit perfect transmission of voicing information for the unmodified stimuli as well as the stimuli with modified transition duration.

For the 12 dB SNR case, the overall information transmitted increases for S1 and S3 by 2% and 19% respectively and decreases for S2 and S4 by 3% and 4% respectively when FTD is doubled. Transmission of place information is seen to increase for S1, S2, and S3 by 13%, 6%, and 33% respectively and decrease for S4 by 5% when FTD is doubled. Transmission of voicing information is seen to decrease for all four subjects as transition duration is increased

For the 6 dB SNR case, the overall information transmitted increases for S1 and S3 by 14% and 26% respectively for 50% FTD increase and then decreases by 7% and 11% respectively for 100% FTD increase. The overall information transmitted increases for S2 by 9% and decreases for S4 by 2% when FTD is doubled. Transmission of place information increases for S1, S2, and S4 by 30%, 15%, and 3% respectively for 50% FTD increase and then decreases by 6%, 2%, and 3% respectively for 100% FTD increase. S3 shows an increase of 33% in the transmission of place information when FTD is doubled. Transmission of voicing information is seen to generally decrease as transition duration is increased.

The relative information transmitted, averaged across the four subjects, for the formant transition duration modification test, are seen in the last row for each feature in Table 7.6. The overall information transmitted decreases by 4% in the no masking noise case and increases by 3% in the 12 dB SNR case when FTD is doubled. The overall information transmitted in the 6 dB SNR case increases by 9% for 50% FTD increase and then decreases by 2% for 100% FTD increase. The transmission of place information decreases by 10% in the no masking noise case and increases by 12% in the 12 dB SNR case when FTD is doubled. The transmission of place information in the 6 dB SNR case increases by 19% for 50% FTD increase and then decreases by 2% for 100% FTD increase. The transmission of voicing information is generally perfect in the no masking noise case and decreases by 27% in the 12 dB SNR case and by 25% in the 6 dB SNR case when FTD is doubled. These results are plotted in Fig. 7.5.

Information transmission analysis results for the <u>voice onset time (VOT) modification</u> test stimuli are summarized in Table 7.7. For the no masking noise case, the overall information transmitted decreases for all the four subjects as VOT is increased. Transmission of place and voicing information is also seen to generally decrease for all the four subjects as VOT is increased.

For the 12 dB SNR case, the overall information transmitted increases for S1 by 10% and decreases for S2, S3, and S4 by 21%, 7%, and 8% respectively when VOT is doubled. Transmission of place information increases for S1 by 13% and decreases for S2, S3, and S4 by 36%, 12%, and 13% respectively when VOT is doubled.

Transmission of voicing information decreases for all the four subjects as VOT is increased.

For the 6 dB SNR case, the overall information transmitted increases for S1 by 2% and decreases for S2, S3, and S4 by 17%, 16%, and 7% respectively when VOT is doubled. Transmission of place information increases for S1 by 10% and decreases for S2, S3, and S4 by 19%, 15%, and 12% respectively when VOT is doubled. Transmission of voicing information decreases for all the four subjects as VOT is increased.

The relative information transmitted, averaged across the four subjects, for the VOT modification test, are seen in the last row for each feature in Table 7.7. The overall information transmitted decreases by 10% in the no masking noise case, 9% in the 12 dB SNR case, and 13% in the 6 dB SNR case when VOT is doubled. Transmission of place information decreases by 4% in the no masking noise case, 13% in the 12 dB SNR case, and 10% in the 6 dB SNR case. Transmission of voicing information decreases by 23% in the no masking noise case, 22% in the 12 dB SNR case, and 32% in the 6 dB SNR case. These results are plotted in Fig. 7.6.

7.4.3 Response Time

The response times for the three tests are given in Table 7.8 for individual subjects, and also averaged across the four subjects.

For the BD modification test, the response times for the four subjects are seen to generally decrease (improve) when burst duration is doubled.

For the FTD modification test all the four subjects show an increase in response time as FTD is increased for the no-noise case. While S1 and S3 show an improvement in response times with increasing FTD for the 12 dB and 6 dB SNR cases, S4 shows a slight increase in response time. The response time for S2 is seen to improve with increasing FTD in the 6 dB SNR case. The response time, averaged across the four subjects, is seen to increase in the no noise case and decrease in the 12 dB SNR case when FTD is increased. In the 6 dB SNR case, the average response time is seen to decrease for 50% FTD increase and remain unaffected thereafter for 100% FTD increase.

For the VOT modification test, the four subjects exhibit a general increase in response times with increasing VOT for the three noise conditions.

The response times, averaged across the four subjects were subjected to the paired t-test and the results are given in Table 7.9. For the burst duration modification test, no significant changes in response times are observed in the no masking noise and 12 dB SNR cases when the burst duration is doubled. However, for the 6 dB SNR case, the response time improves fairly significantly $(p<0.2)$ when burst duration is doubled.

For the formant transition duration modification test, no significant change in response times are observed in both the no masking noise and 12 dB SNR cases when the transition duration is increased. However, in the 6 dB SNR case, a slight improvement $(p<0.5)$ in response time is observed when the transition duration is doubled.

For the voice onset time modification test, the response time is seen to increase (p<0.2) in the no masking noise case when VOT is doubled. A similar situation is observed for the 6 dB SNR case too.

7.5 CONCLUSIONS

From the recognition scores and information transmission analysis for the no-noise and 12 dB SNR cases, one of the four subjects shows an appreciable increase in performance when burst duration is doubled. However, at 6 dB SNR, three of the subjects show an improvement in performance. This seems to suggest that some subjects may find burst duration modification beneficial at lower SNRs.

For the FTD modification test, as SNR is reduced, three of the subjects show a general increase in recognition scores as FTD is increased to 50% and then a decrease in scores as FTD is increased to 100%. For the the fourth subject, a falling trend in scores is observed as FTD is increased to 50% and then to 100%. However, information transmission analysis reveals a general increase in overall information transmitted as FTD is increased to 50% which then decreases when FTD is increased to 100%. This suggests that for the stimuli considered here an FTD increase upto 50% would be beneficial beyond which the scores and overall information transmission would generally decrease.

For the VOT modification test, the recognition scores as well as information transmission analysis show a slight benefit in increasing VOT for one subject. As SNR is reduced, the amount of benefit accruing from increased VOT for that subject is seen to decrease. However, for the other three subjects the performance is seen to generally decrease with increasing VOT for all the three noise cases.

The average response time is seen to generally decrease (improve) with increase in burst duration and formant transition duration. However, with increasing VOT, the average response time is seen to generally increase.

The above results seem to suggest that of the three acoustic segments considered here that contribute to increased consonant duration in clear speech, formant transition duration yields the most positive results. An FTD increase upto 50% is seen to improve the performance for all the three subjects. At lower SNRs, this amount of FTD increase may be combined with burst duration modification and expected to yield better performance. However, voice onset time does not appear to be a suitable candidate for modification because of the reduction in performance seen with increasing VOT.

TABLE 7.1 Burst duration modification test: Consonant recognition scores, for different amounts of burst duration (BD) increase and SNRs, for the four subjects. Last row gives the recognition scores and standard deviations, averaged across the four subjects.

Sub- ject	Consonant recognition scores (%)					
	No masking noise		SNR = 12 dB		SNR = 6 dB	
	BD increase		BD increase		BD increase	
	0%	100%	0%	100%	0%	100%
S1	98	94	91	89	73	77
S2	99	96	82	92	73	80
S3	84	92	83	84	69	77
S4	100	100	93	89	86	79
Avg (SD)	96 (7.3)	96 (3.3)	88 (5.4)	89 (4.1)	77 (7.2)	78 (1.7)

TABLE 7.2 Formant transition duration modification test: Consonant recognition scores, for different amounts of formant transition duration (FTD) increase and SNRs, for the four subjects. Last row gives the recognition scores and standard deviations, averaged across the four subjects.

Sub- ject	Consonant recognition scores (%)								
	No masking noise			SNR = 12 dB			SNR = 6 dB		
	FTD increase			FTD increase			FTD increase		
	0%	50%	100%	0%	50%	100%	0%	50%	100%
S1	99	97	96	87	87	84	70	86	80
S2	99	90	87	88	79	73	69	73	64
S3	86	80	84	76	82	89	71	87	82
S4	100	99	100	89	83	78	84	73	73
Avg (SD)	96 (6.9)	92 (8.4)	93 (7.4)	85 (6.1)	83 (3.2)	81 (6.8)	75 (7.1)	79 (7.4)	75 (8.0)

TABLE 7.3 Voice onset time modification test: Consonant recognition scores for different amounts of voice onset time (VOT) increase and SNRs, for the four subjects. Last row gives the recognition scores and standard deviations, averaged across the four subjects.

Subject	Consonant recognition scores (%)								
	No masking noise			SNR = 12 dB			SNR = 6 dB		
	VOT increase			VOT increase			VOT increase		
	0%	50%	100%	0%	50%	100%	0%	50%	100%
S1	99	96	93	86	87	92	73	78	81
S2	93	87	84	92	74	70	80	79	59
S3	83	91	81	80	77	73	76	69	69
S4	100	99	97	90	83	87	81	80	81
Avg (SD)	**95** (7.6)	**94** (5.2)	**91** (7.5)	**88** (5.4)	**81** (5.7)	**82** (10.6)	**79** (3.8)	**77** (5.1)	**75** (10.7)

TABLE 7.4 Paired t-test significance levels of consonant recognition scores of Tables 7.1 to 7.3 for the three tests, averaged across the four subjects. (BD: Burst duration; FTD: Formant transition duration; VOT: Voice onset time). N=4.

Test	SNR (dB)	Modified parameter	mean	SD	Test of difference (two-tailed)	
					t	p
		BD increase (%)				
BD	No masking noise	0	96.2	7.3		
		100	96.4	3.3	0.05	NS
	12	0	88.4	5.4		
		100	89.1	4.1	0.21	NS
	6	0	77.3	7.2		
		100	78.4	1.7	0.28	NS
		FTD increase (%)				
FTD	No masking noise	0	96.4	6.9		
		50	92.4	8.4	0.74	<0.5
		100	92.8	7.4	0.71	NS
	12	0	85.2	6.1		
		50	82.6	3.2	0.76	<0.5
		100	80.7	6.8	0.99	<0.4
	6	0	75.0	7.1		
		50	78.8	7.4	0.74	<0.5
		100	74.5	8.0	0.09	NS
		VOT increase (%)				
VOT	No masking noise	0	95.4	7.6		
		50	94.4	5.2	0.22	NS
		100	91.0	7.5	0.82	<0.5
	12	0	87.7	5.4		
		50	81.0	5.7	1.71	<0.2
		100	82.1	10.6	0.94	<0.4
	6	0	78.5	3.8		
		50	77.3	5.1	0.38	NS
		100	74.6	10.7	0.69	NS

TABLE 7.5 Burst duration (BD) modification test: Information transmission analysis for the stops /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/, for the four subjects. Last row for each feature gives the relative information transmitted, averaged across the four subjects.

Feature	Subject	Relative information transmitted (%)					
		SNR = No noise		SNR = 12 dB		SNR = 6 dB	
		BD increase		BD increase		BD increase	
		0%	100%	0%	100%	0%	100%
Overall	S1	96	92	86	82	70	71
	S2	98	92	77	90	64	70
	S3	87	89	79	75	64	73
	S4	100	100	85	83	76	70
	Avg	**94**	**92**	**80**	**80**	**66**	**68**
Place	S1	92	100	76	71	40	48
	S2	96	100	57	79	43	57
	S3	67	81	66	63	44	56
	S4	100	100	78	74	62	55
	Avg	**87**	**94**	**68**	**70**	**47**	**52**
Voicing	S1	100	70	100	82	100	87
	S2	100	74	92	100	92	79
	S3	100	92	100	87	78	92
	S4	100	100	91	96	96	85
	Avg	**100**	**85**	**94**	**91**	**90**	**85**

TABLE 7.6 Formant transition duration (FTD) modification test: Information transmission analysis for the stops /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/, for the four subjects. Last row for each feature gives the relative information transmitted, averaged across the four subjects.

Feature	Sub-ject	Relative information transmitted (%)								
		SNR = No noise			SNR = 12 dB			SNR = 6 dB		
		FTD increase			FTD increase			FTD increase		
		0%	50%	100%	0%	50%	100%	0%	50%	100%
Overall	S1	98	94	93	79	82	81	62	76	69
	S2	98	89	86	82	78	79	66	71	75
	S3	85	85	83	67	83	86	62	88	77
	S4	100	97	100	83	75	79	76	73	74
	Avg	**93**	**89**	**89**	**76**	**77**	**79**	**63**	**72**	**70**
Place	S1	96	92	85	62	70	75	31	61	55
	S2	96	75	70	68	66	74	42	57	55
	S3	65	59	61	44	66	78	39	70	72
	S4	100	95	100	73	71	68	61	64	61
	Avg	**87**	**78**	**77**	**61**	**68**	**73**	**41**	**60**	**58**
Voicing	S1	100	92	100	92	92	70	87	87	70
	S2	100	100	100	92	78	48	87	74	53
	S3	100	100	100	92	92	84	82	100	67
	S4	100	100	100	95	66	68	95	65	63
	Avg	**100**	**98**	**100**	**92**	**78**	**65**	**87**	**76**	**62**

TABLE 7.7 Voice onset time (VOT) modification test: Information transmission analysis for the stops /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/, for the four subjects. Last row for each feature gives the relative information transmitted, averaged across the four subjects.

Feature	Sub-ject	Relative information transmitted (%)								
		SNR = No noise			SNR = 12 dB			SNR = 6 dB		
		VOT increase			VOT increase			VOT increase		
		0%	50%	100%	0%	50%	100%	0%	50%	100%
Overall	S1	98	91	90	78	86	88	70	70	72
	S2	89	82	76	86	67	65	75	67	58
	S3	87	88	78	71	67	64	74	55	58
	S4	100	97	94	85	74	77	75	71	68
	Avg	**92**	**88**	**82**	**79**	**70**	**70**	**71**	**64**	**58**
Place	S1	96	88	88	63	68	76	40	46	50
	S2	78	67	75	76	50	40	56	51	37
	S3	66	79	65	59	49	47	50	34	35
	S4	100	100	95	75	56	62	58	48	46
	Avg	**84**	**84**	**80**	**68**	**53**	**55**	**49**	**44**	**39**
Voicing	S1	100	92	87	92	92	87	100	92	87
	S2	100	92	67	100	74	61	92	79	49
	S3	100	92	67	79	74	41	100	57	36
	S4	100	91	84	100	93	91	96	96	82
	Avg	**100**	**92**	**77**	**93**	**85**	**71**	**96**	**82**	**64**

TABLE 7.8 Response times in sec (and standard deviations) for the four subjects. (BD: Burst duration; FTD: Formant transition duration; VOT: Voice onset time).

(a) Burst duration modification test:

Subject	Response time (sec)					
	SNR = No noise		SNR = 12 dB		SNR = 6 dB	
	BD increase		BD increase		BD increase	
	0%	100%	0%	100%	0%	100%
S1	1.36 (0.05)	1.51 (0.12)	1.86 (0.23)	1.61 (0.16)	2.15 (0.24)	1.82 (0.23)
S2	1.66 (0.14)	1.48 (0.09)	2.01 (0.07)	1.92 (0.08)	2.35 (0.46)	1.87 (0.29)
S3	2.49 (0.08)	2.41 (0.06)	2.58 (0.28)	2.87 (0.35)	3.40 (1.26)	2.60 (0.20)
S4	1.59 (0.03)	1.52 (0.11)	2.26 (0.32)	2.15 (0.31)	2.56 (0.25)	2.46 (0.32)
Avg	**1.78** (0.49)	**1.73** (0.45)	**2.18** (0.32)	**2.14** (0.54)	**2.62** (0.55)	**2.19** (0.40)

(b) Formant transition duration modification test:

Subject	Response time (sec)								
	SNR = No noise			SNR = 12 dB			SNR = 6 dB		
	FTD increase			FTD increase			FTD increase		
	0%	50%	100%	0%	50%	100%	0%	50%	100%
S1	1.39 (0.28)	1.35 (0.04)	1.50 (0.13)	1.84 (0.03)	1.78 (0.06)	1.65 (0.13)	2.01 (0.17)	1.77 (0.08)	1.93 (0.23)
S2	1.57 (0.24)	1.65 (0.19)	1.77 (0.30)	1.65 (0.10)	1.89 (0.14)	1.78 (0.17)	2.25 (0.58)	1.90 (0.21)	1.94 (0.28)
S3	2.55 (0.20)	2.56 (0.39)	2.62 (0.21)	2.86 (0.27)	2.40 (0.08)	2.30 (0.05)	2.86 (0.36)	2.60 (0.15)	2.40 (0.05)
S4	1.44 (0.04)	1.57 (0.08)	1.66 (0.10)	2.12 (0.36)	2.24 (0.46)	2.40 (0.39)	2.32 (0.38)	2.41 (0.60)	2.41 (0.25)
Avg	**1.74** (0.55)	**1.78** (0.53)	**1.89** (0.50)	**2.12** (0.53)	**2.08** (0.29)	**2.03** (0.37)	**2.36** (0.36)	**2.17** (0.40)	**2.17** (0.27)

Table 7.8 cont...

...Table 7.8 cont

(c) Voice onset time modification test:

Subject	Response time (sec)								
	SNR = No noise			SNR = 12 dB			SNR = 6 dB		
	VOT increase			VOT increase			VOT increase		
	0%	50%	100%	0%	50%	100%	0%	50%	100%
S1	1.39 (0.28)	1.83 (0.26)	1.79 (0.03)	1.84 (0.09)	2.10 (0.04)	1.83 (0.13)	2.15 (0.24)	2.10 (0.36)	2.06 (0.26)
S2	1.72 (0.18)	2.54 (0.17)	2.60 (0.62)	2.49 (0.56)	2.45 (0.32)	2.28 (0.29)	2.22 (0.08)	2.24 (0.22)	2.68 (0.32)
S3	2.17 (0.39)	2.36 (0.14)	2.54 (0.16)	2.58 (0.13)	2.77 (0.12)	2.74 (0.45)	2.49 (0.08)	2.61 (0.12)	2.51 (0.14)
S4	1.32 (0.05)	1.57 (0.05)	1.88 (0.33)	1.88 (0.23)	2.27 (0.29)	2.21 (0.20)	2.07 (0.16)	2.34 (0.31)	2.32 (0.21)
Avg	1.65 (0.39)	2.07 (0.45)	2.20 (0.43)	2.20 (0.39)	2.40 (0.29)	2.27 (0.37)	2.23 (0.19)	2.32 (0.22)	2.39 (0.27)

TABLE 7.9 Paired t-test significance levels of response times (in sec) of Table 7.8 for the three tests, averaged across the four subjects. (BD: Burst duration; FTD: Formant transition duration; VOT: Voice onset time). N=4.

Test	SNR (dB)	Modified parameter	mean	SD	Test of difference (two-tailed)	
					t	p
		BD increase (%)				
BD	No masking noise	0	1.78	0.49		
		100	1.73	0.45	0.15	NS
	12	0	2.18	0.32		
		100	2.14	0.54	0.13	NS
	6	0	2.62	0.55		
		100	2.19	0.40	1.44	<0.2
		FTD increase (%)				
FTD	No masking noise	0	1.74	0.55		
		50	1.78	0.53	0.10	NS
		100	1.89	0.50	0.40	NS
	12	0	2.12	0.53		
		50	2.08	0.29	0.13	NS
		100	2.03	0.37	0.28	NS
	6	0	2.36	0.36		
		50	2.17	0.40	0.71	NS
		100	2.17	0.27	0.84	<0.5
		VOT increase (%)				
VOT	No masking noise	0	1.65	0.39		
		50	2.07	0.45	1.41	<0.4
		100	2.20	0.43	1.89	<0.2
	12	0	2.20	0.39		
		50	2.40	0.29	0.82	<0.5
		100	2.27	0.37	0.26	NS
	6	0	2.23	0.19		
		50	2.32	0.22	0.62	NS
		100	2.39	0.27	0.97	<0.4

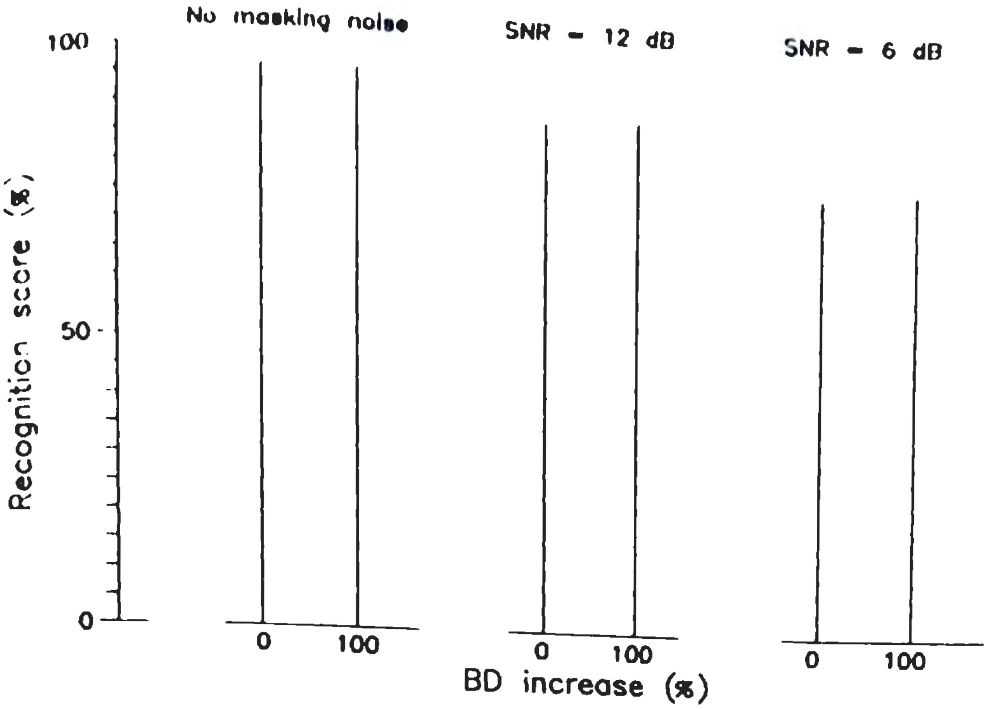


FIG. 7.1 BD modification test: Percentage recognition score versus burst duration increase averaged across the four subjects.

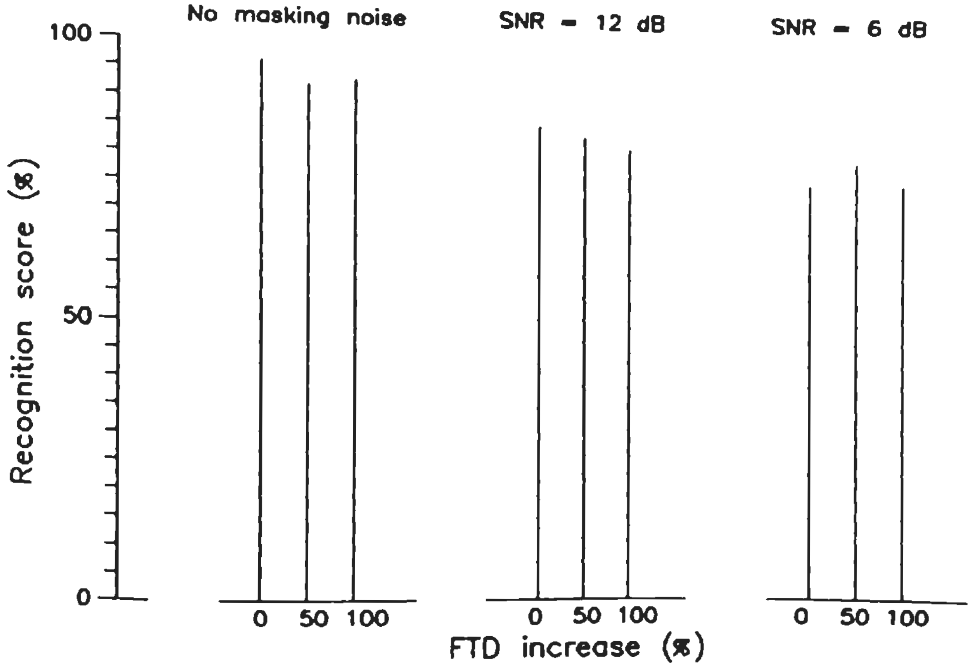


FIG. 7.2 FTD modification test: Percentage recognition score versus formant transition duration increase averaged across the four subjects.

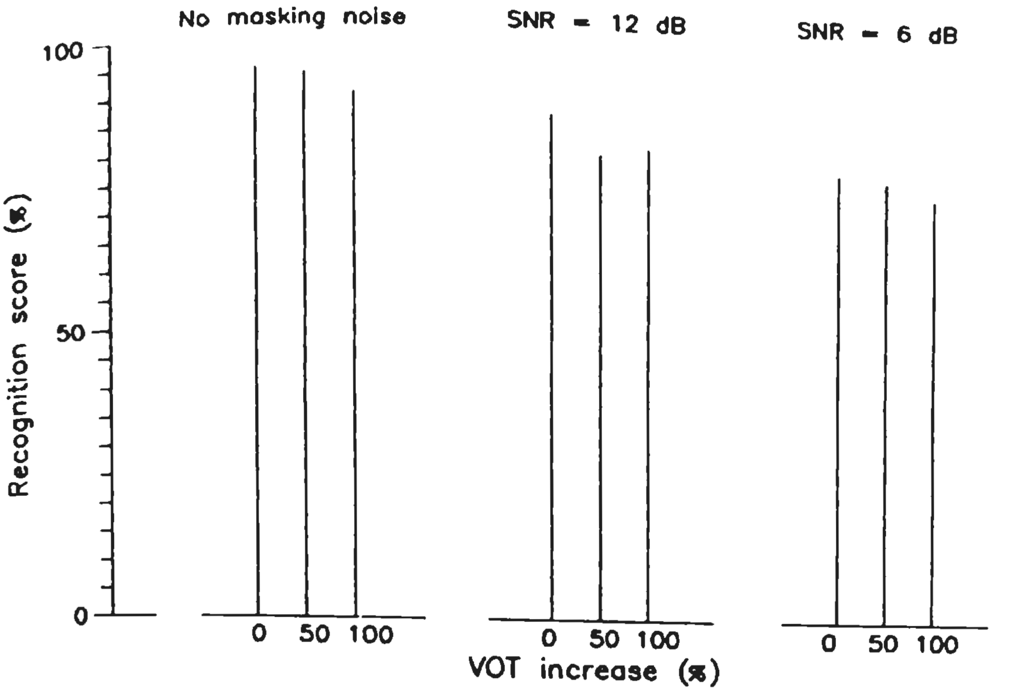


FIG. 7.3 VOT modification test: Percentage recognition score versus voice onset time increase averaged across the four subjects.

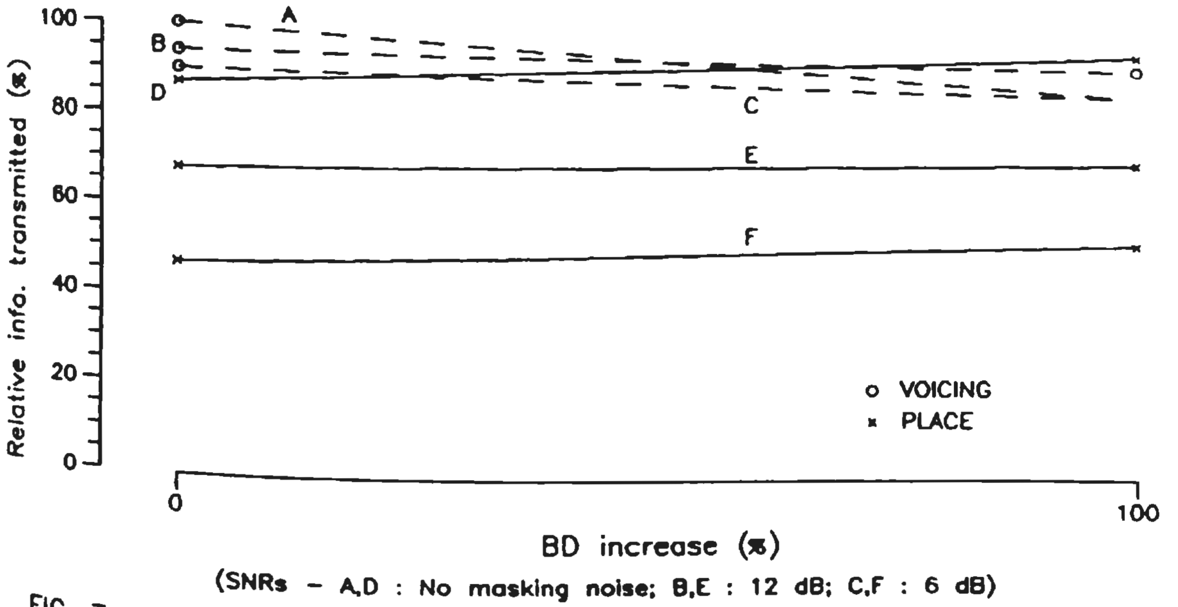
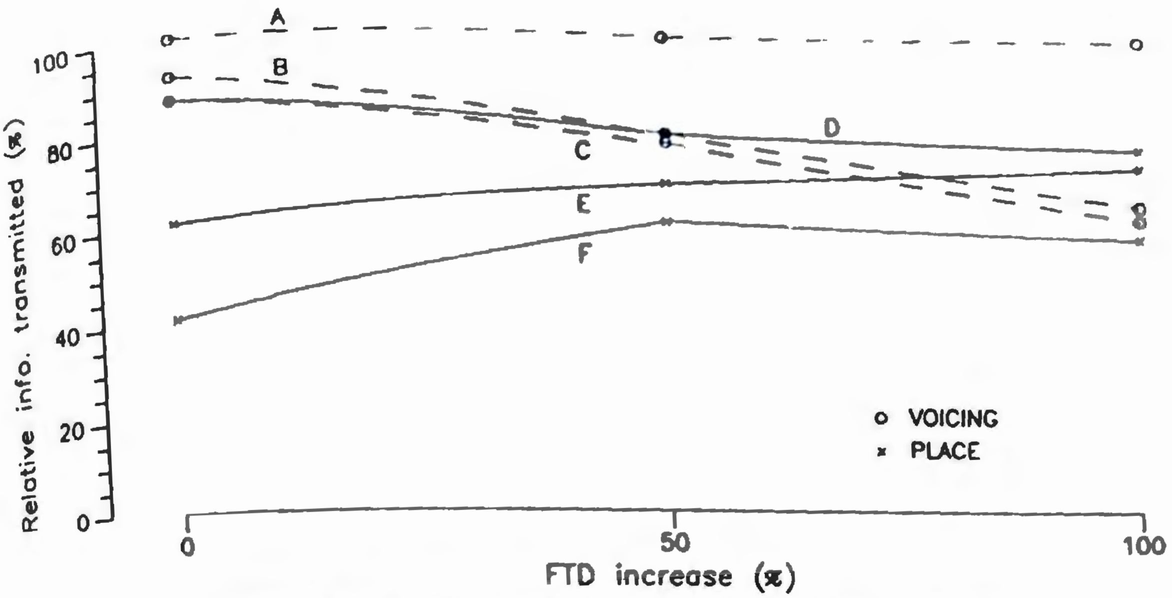
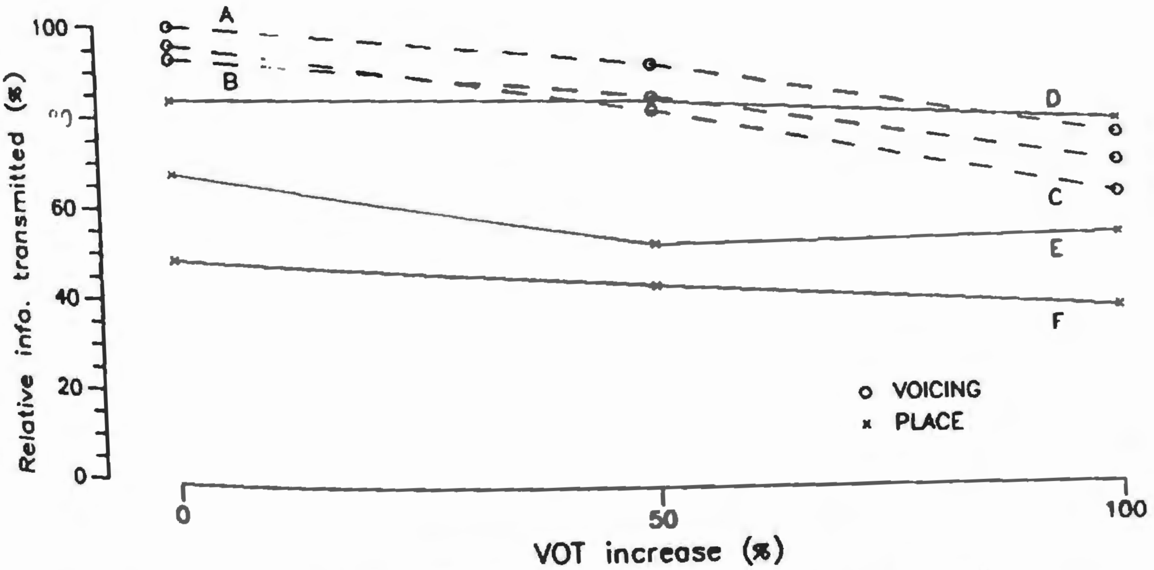


FIG. 7.4 BD modification test: The relative information transmitted about place and voicing features plotted as a function of burst duration increase for different SNRs. (Data points are connected by straight line)



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 7.5 FTD modification test: The relative information transmitted about place and voicing features plotted as a function of formant transition duration increase for different SNRs. (Data points are connected by straight line)



(SNRs – A,D : No masking noise; B,E : 12 dB; C,F : 6 dB)

FIG. 7.6 VOT modification test: The relative information transmitted about place and voicing features plotted as a function of voice onset time increase for different SNRs. (Data points are connected by straight line)

CHAPTER 8

CONCLUSIONS

8.1 INTRODUCTION

A promising scheme for enhancing the speech signal is based on studies of speaking clearly for the hearing impaired. Studies of the differences between "clear" speech and "conversational" speech have identified certain consistent acoustic modifications of the speech signal in clear speech. Two characteristics of clear (intelligible) speech, namely consonant-to-vowel (C/V) intensity ratio and consonant duration, have been chosen for evaluation in the present study.

A few studies reported in the literature have evaluated the intelligibility of natural speech artificially transformed to clear speech by altering either one or both of the above acoustic parameters. While all the studies reported improvement in C/V intensity ratio modification, the effect of increasing consonantal duration has been equivocal. This may have been because the acoustic subsegments associated with consonant phonemes increase in a nonuniform manner and such changes cannot be aptly simulated by merely replicating portions of the phonemic subsegments in order to extend consonantal duration. What is called for is some method whereby the effects, on perception, of increasing the duration of each acoustic segment independently can be monitored. This can be achieved by using synthesized speech material. In this study, a modified form of the software based formant synthesizer, earlier developed by Klatt [56] was used to generate the test stimuli. Various modifications can be introduced by appropriately altering the synthesis parameters. Thus emphasis on various acoustic segments can be altered independently.

It was of interest to study which particular phonemic features would be aided in perception by the processing and whether it would have any adverse effect on the perception of the accompanying vowels. Hence, it was decided to conduct experiments on closed-set CV and VC syllables.

Realizing the attendant difficulties in carrying out such a study on subjects with sensorineural hearing impairment, the experiments were carried out on normal-hearing subjects with hearing impairment being simulated by presenting the sounds in a background of flat-spectrum noise. This simulates a moderate sensorineural loss, the extent of which can be controlled by varying the levels of background noise.

A PC-based test environment was created for: (i) test administration, (ii) acquisition of data, (iii) data analysis and various modes of graphical presentation - all of which should be of use in future perception studies.

The results obtained from the experiments on C/V intensity ratio modification and consonant duration modification are summarized in the following sections.

8.2 EXPERIMENTS ON C/V INTENSITY RATIO MODIFICATION

The experiments involving consonant-to-vowel intensity ratio (CVR) modification were divided into four test sets: CV9, VC9, CV6 and VC6. The CV9 and VC9 tests involved the three unvoiced stop consonants /p, t, k/ in the CV and VC contexts respectively of the cardinal vowels /a, i, u/. The CV6 and VC6 tests involved the six stop consonants /p, t, k, b, d, g/ in the CV and VC contexts respectively of the vowel /a/. Five versions of each stimulus were synthesized, one with no modification and the other four with CVR modification of +3, +6, +9, and +12 dB. To simulate hearing impairment, each stimulus was mixed with synthesized broadband noise under three SNR conditions: no masking noise, and masking noise with 12 dB SNR and 6 dB SNR.

The consonant recognition scores obtained for all the four tests show no significant changes when the C/V ratio is increased in the no masking noise case. However, in the presence of masking broadband noise, the recognition scores are seen to improve significantly with increasing CVR for all the tests. Another important finding is that the scores for stops in the syllable-final position (VC) are seen to be higher than those for stops in the syllable-initial (CV) position for all the CVRs. These results have also been borne out by statistical analysis.

The pattern of confusions in the identification of place of articulation remains generally unaffected for all the CVR modifications in all the four tests. Since increase in CVR has been seen to be more effective in bringing down the overall level of confusions in the VC context as compared to that in the CV context, it appears that increasing CVR suppresses forward masking of the consonant by the vowel more effectively than backward masking.

Increasing CVR can sometimes result in vowel confusions as seen from the /ip/-/up/ confusions which increased in severity for higher CVR modifications. Thus, the possibility of vowel confusions is another factor that could set a limit on the amount of CVR modification that may be used. In experiments with closed-set stimuli, this limit as well as the specific vowel confusions may depend upon the test set involved.

Information transmission analysis results for the CV9 and VC9 tests show that both overall information transmitted as well as transmission of consonant feature increases appreciably with increasing CVR for all the SNRs. However, the overall information transmitted as well as transmission of place feature is seen to be superior in the VC context as compared to those in the CV context. The information transmission of vowel feature for the two tests is seen to be near-perfect for the no masking noise and 12 dB SNR conditions with a slight decrease for the 6 dB SNR condition.

Information transmission analysis results for the CV6 and VC6 tests show near-perfect transmission of overall information as well as information transmitted with respect to place and voicing features for the no masking noise case. For the 12 dB and 6 dB SNR cases, the information transmitted with respect to all the three features is found to increase appreciably with increase in CVR modification.

The results of the information transmission analysis for CV6 and VC6 tests further reveal that the information transmitted about place feature is superior in the VC context over that in the CV context. However, the transmission of voicing information is superior in the CV context over that in the VC context.

The average response times for all the four tests do not appear to show any deterioration and in some cases appear to indicate a quicker response by the subjects for increasing values of CVR modification. This has also been borne out by statistical analysis of the results.

These set of experiments were done to study the effect of increasing CVR on stop consonants from three different view-points: recognition score, amount of information transmitted on the basis of feature classification, and the average response time. As seen from the above study, increasing CVR had a positive impact in all the cases. A CVR modification of upto about 10 dB may be used without any adverse effect on vowel recognition for the test stimuli considered here.

8.3 EXPERIMENTS ON CONSONANT DURATION MODIFICATION

The experiments involving consonant duration modification were performed to study the effect on perception of altering the duration of each of the acoustic segments that constitute the phoneme. Stop consonants /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/ were chosen as the stimuli to separately consider the effect of modifying the following: burst duration (BD), formant transition duration (FTD), and voice onset time (VOT).

For the BD modification test in the no masking noise and 12 dB SNR cases , one of the four subjects shows an appreciable increase in performance when burst duration is doubled. However, at 6 dB SNR, three of the subjects show an improvement in performance. This suggests that some subjects may find burst duration modification beneficial at lower SNRs.

For the FTD modification test, as SNR is reduced, three of the subjects show a general increase in recognition scores as FTD is increased to 50% and then a decrease in scores as FTD is increased to 100%. For the the fourth subject, a falling trend in scores is observed as FTD is increased to 50% and then to 100%. However, information transmission analysis reveals a general increase in overall information transmitted as FTD is increased to 50% which then decreases when FTD is increased to 100%. This suggests that for the stimuli considered here an FTD increase upto 50% would be beneficial beyond which the scores and overall information transmission would generally decrease.

For the VOT modification test, the recognition scores as well as information transmission analysis show a slight benefit in increasing VOT for one subject. As SNR is reduced, the amount of benefit accruing from increased VOT for that subject is seen to decrease. However, for the other three subjects the performance is seen to generally decrease with increasing VOT for all the three noise cases.

The average response time is seen to generally decrease (improve) with increase in burst duration and formant transition duration. However, with increasing VOT, the average response time is seen to generally increase.

The above results have also been verified by statistical analysis. Thus, of the three acoustic segments considered here that contribute to increased consonant duration in clear speech, formant transition duration yields the most positive results. An FTD increase upto 50% is seen to improve the performance for all the three subjects. At lower SNRs, this amount of FTD increase may be combined with burst duration modification and expected to yield better performance. However, voice onset time does not appear to be a suitable candidate for modification.

8.4 C/V INTENSITY RATIO (CVR) MODIFICATION VERSUS CONSONANT DURATION (CD) MODIFICATION

The above results have shown that increasing CVR does improve recognition scores. Some vowel confusions are observed in the VC context at higher CVRs, but as long as CVR modification is restricted to about 10 dB, there is no adverse effect on the recognition of vowels. The information transmitted on the basis of feature classification, as well as the average response times were found to improve with increasing CVR.

For the CD modification experiments, the acoustic segments that constitute the consonant phonemes were altered in duration separately to study their individual effects. The results suggest that at higher noise levels, a formant transition duration modification of upto about 50% may be combined with burst duration modification and expected to yield better performance. Voice onset time (VOT) does not appear to be a suitable parameter for modification as performance decreased with increasing VOT.

8.5 SUGGESTIONS FOR FUTURE WORK

As summarized in the previous section, the present study has established that consonant-to-vowel intensity ratio modification and, to a lesser extent, modification in formant transition duration would improve speech perception. This study was done using nonsense CV and VC syllables as test material and subjects with simulated sensorineural impairment. In order to further evaluate the effectiveness of this approach, work should be carried out on building a speech processor and conducting experiments on subjects with hearing impairment.

In the earlier studies reported, the modifications have been done on natural speech by visual inspection of the digitized waveform, intensity envelope, and spectrogram. C/V ratio was altered by modifying the intensity envelope. The consonant duration was increased by merely replicating small time sections within the consonantal subsegments. This cannot be expected to reflect the actual changes associated with acoustic segments. In this study, a speech synthesizer has been used for generating the test stimuli, and the parameters were modified during the synthesis stage.

In building a speech processor, an analysis/synthesis approach would possibly have to be adopted in which the subsegment boundaries can be identified and the subsegments resynthesized after appropriate modifications. It is obvious that this analysis/synthesis approach would require a processing time delay extending over several subsegment durations. It has been previously shown that for speech processing for cochlear prostheses, delays of upto about 120 ms in the speech processing and stimulus encoding should not interfere with the benefits of auditory signal in audiovisual comprehension of connected speech [78]. Hence, in the present case, the segmentation, feature classification, and speech enhancement processes should be completed within this time.

APPENDIX A
SPECTROGRAPHIC ANALYSIS OF SPEECH WAVEFORMS

A.1 INTRODUCTION

The analysis of speech signals involve measurements of (a) temporal features, (b) spectral features, and (c) parameters of some model of the speech production system [58]. The time varying spectral characteristics of speech are traditionally displayed as "spectrogram", which is a two-dimensional pattern with the horizontal and vertical axes corresponding to time and frequency respectively and the darkness of the pattern representing the signal energy [59]. A pascal program SPG was developed for displaying the spectral content of dynamic signals like those associated with speech, underwater sounds, and biomedical phenomena [60].

A.2 SPECTROGRAPHIC ANALYSIS

Among the various methods of spectrogram generation, one convenient method is to compute the short-time Fourier transform (STFT) [58] from the sampled waveform as given by:

$$X(n,k) = \sum_{m=-(N/2)}^{(N/2)-1} w(m).x(n-m).\ e^{-j2\pi km/N} \qquad (A.1)$$

where n represents the discrete-time axis, k the discrete frequency axis, and N the discrete Fourier transform (DFT) size. The spectrogram is obtained by displaying the STFT magnitude. The analysis window used, $w(m)$, is an L-point Hamming window. The DFT size, N, used is 512 points. In order to have the same number of spectral samples for different values of L, the sequence of length L is padded with zeroes to make it of length N before performing the DFT.

Traditional spectrograms have either a good time resolution or a good frequency resolution, both of which may exhibit certain characteristic features. The choice of window duration trades off time and frequency resolution [96]. In speech analysis, wideband and narrowband spectrograms are obtained using filter banwidths of 300 Hz and 45 Hz respectively. With speech digitized at 10 K samples/s, and with the use of Hamming window in the analysis, the corresponding window lengths are about 4 ms (L = 43 samples) and about 30 ms (L = 289 samples) respectively.

A.3 COMBINED SPECTROGRAM

Methods have been reported for obtaining a spectrogram-like representation with good time and frequency resolution available in a single display simultaneously [97]. Most of these methods are computationally intensive and the resulting displays are also difficult to interpret. In one of the simpler methods for preserving the features of both wideband and narrowband spectrograms [98], a "combined" spectrogram X_{cb} is obtained by evaluating the geometric mean of the wideband spectrogram X_{wb} and narrowband spectrogram X_{nb}. The geometric mean operation preserves the lighter levels of each of the two spectrograms and hence both the horizontal and vertical features remain visible in the combined spectrogram.

A.4 THE SPECTROGRAM PROGRAM

In addition to the facility for displaying spectrograms with different spectral resolutions, the program SPG also incorporates combined spectrogram display [61]. The spectrogram is displayed by using a VGA card with 16 simultaneous colours and 640x480 screen resolution and a monochrome monitor. The spectrogram is 500 pixels wide, allowing a display of 500 overlapping frames, independent of segment length. It is 256 pixels high. The display also includes a grey-scale plot, vertical axis calibration, time-waveform, cursor readouts, and user directives.

The digitized data file is created using a data acquisition card. The first entry in the file should be the number of samples and this is followed by the signal samples in consecutive lines. The time-waveform is first displayed and the user can select the segment of interest using cursors. The program windows and then pre-emphasizes the selected segment to lower the dynamic range requirement. The STFT is calculated via the fast Fourier transform (FFT). The log magnitude in dB is computed and the spectral information for the frame is displayed by 256 vertical pixels above the trailing edge of the time window. The above procedure is repeated for a new cross-section within the selected segment by shifting the window appropriately. The amount of this shift is obtained such that the spectral information corresponding to 500 cross-sections is ultimately displayed. The display shows log magnitude as a function of frequency (y-axis) for the time-duration (x-axis) of the waveform. The mapping shows high spectral magnitude as black, intermediate magnitude levels in shades of grey and absense of significant magnitude as white. The program allows user set values for maximum and minimum magnitudes so that the dynamic range of the display can be adjusted. Time and frequency cursors are provided whereby readouts of the STFT log magnitude versus frequency at any selected time position can be obtained. In addition, the formant tracks obtained from the display for natural utterances can be used as supplementary information for generating the corresponding synthetic stimuli.

The spectrographic analysis has been subsequently implemented by Baragi & Prasad [102, 103] using a DSP board with TI/TMS 320C25 processor interfaced to the PC bus. The signal can be acquired using the A/D converter of the DSP board or previously digitized signal files can be used. The FFT analysis part is handled by the DSP board while the PC hardware is used for user interface and display operations. This has significantly improved the analysis speed. This program also provides facility for capturing the display screen as bit map for storage and printing on a laser printer after its conversion to postscript format. Hardcopies of the spectrograms in this study were obtained using this system.

Appendix B

PARAMETER TRACKS AND SPECTROGRAMS OF SAMPLE TEST STIMULI

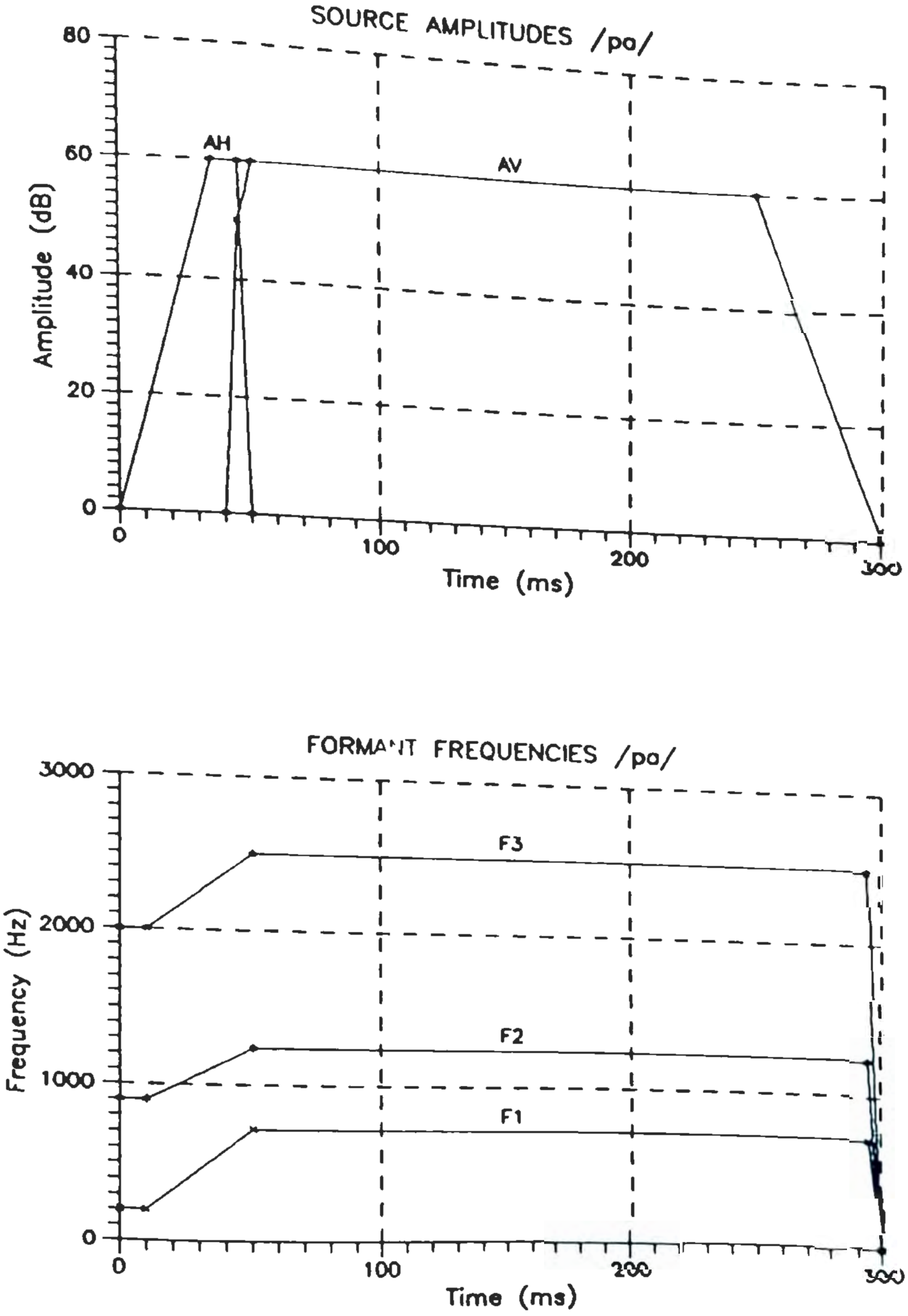
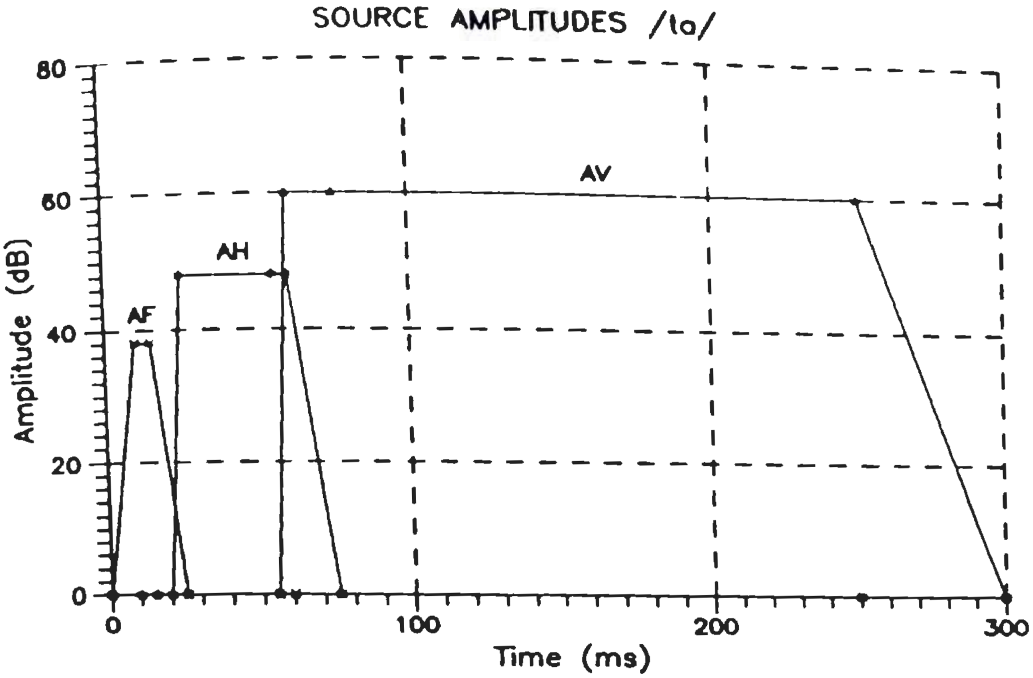


FIG. B.1 Parameter tracks for /pa/.



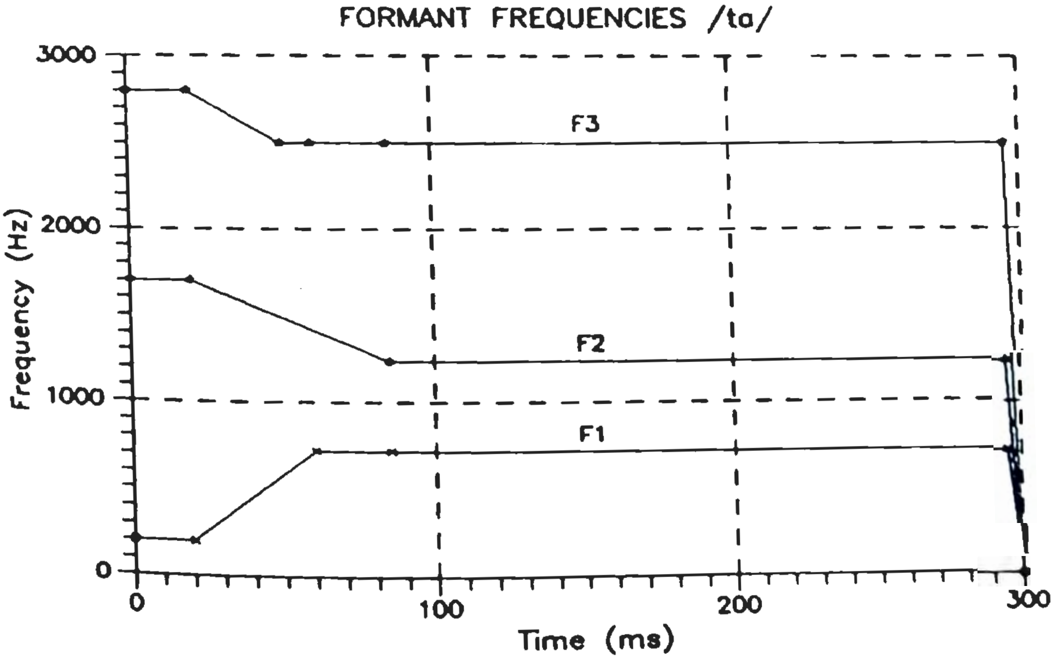


FIG. B.2 Parameter tracks for /ta/.

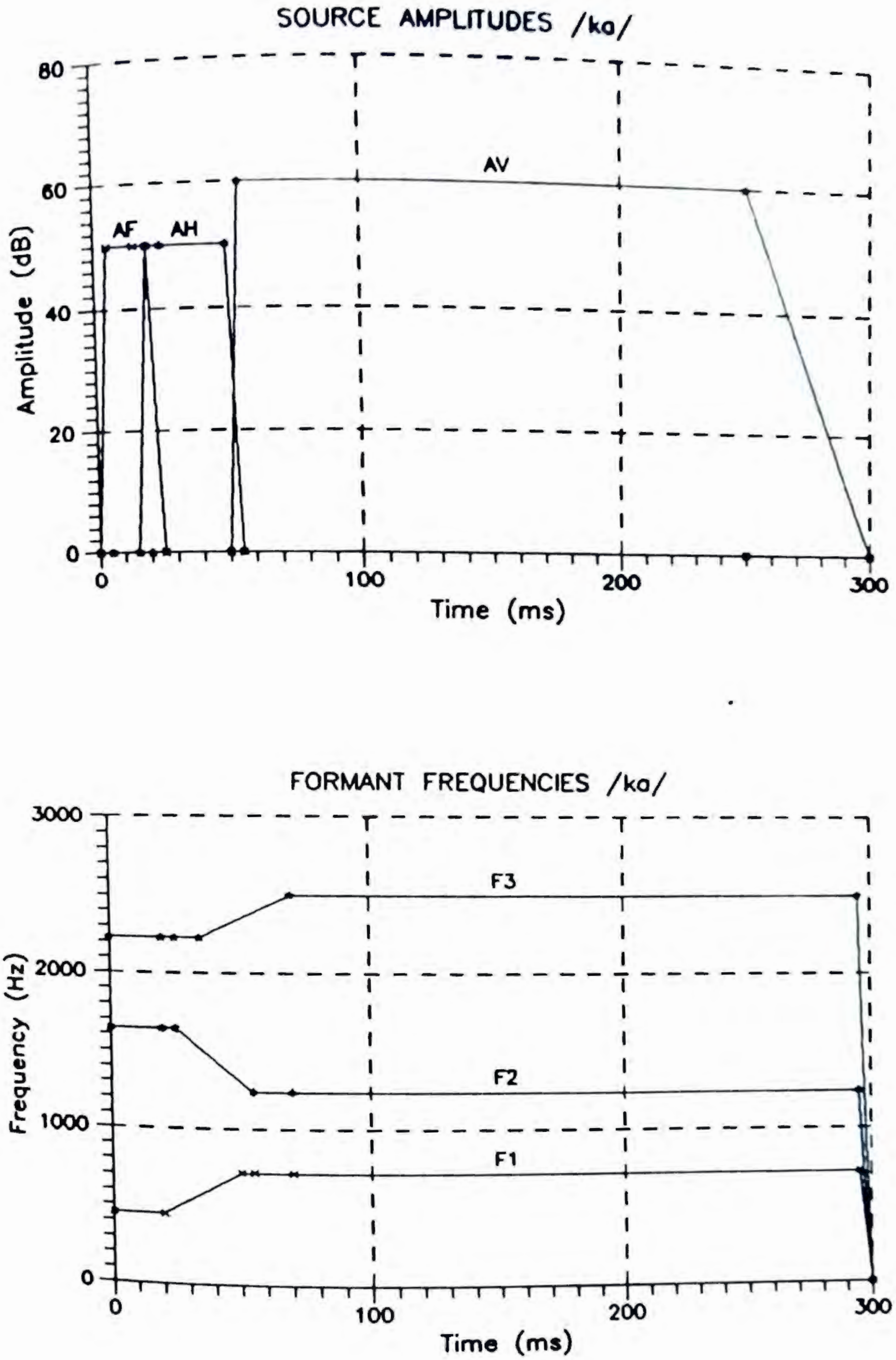


FIG. B.3 Parameter tracks for /ka/.

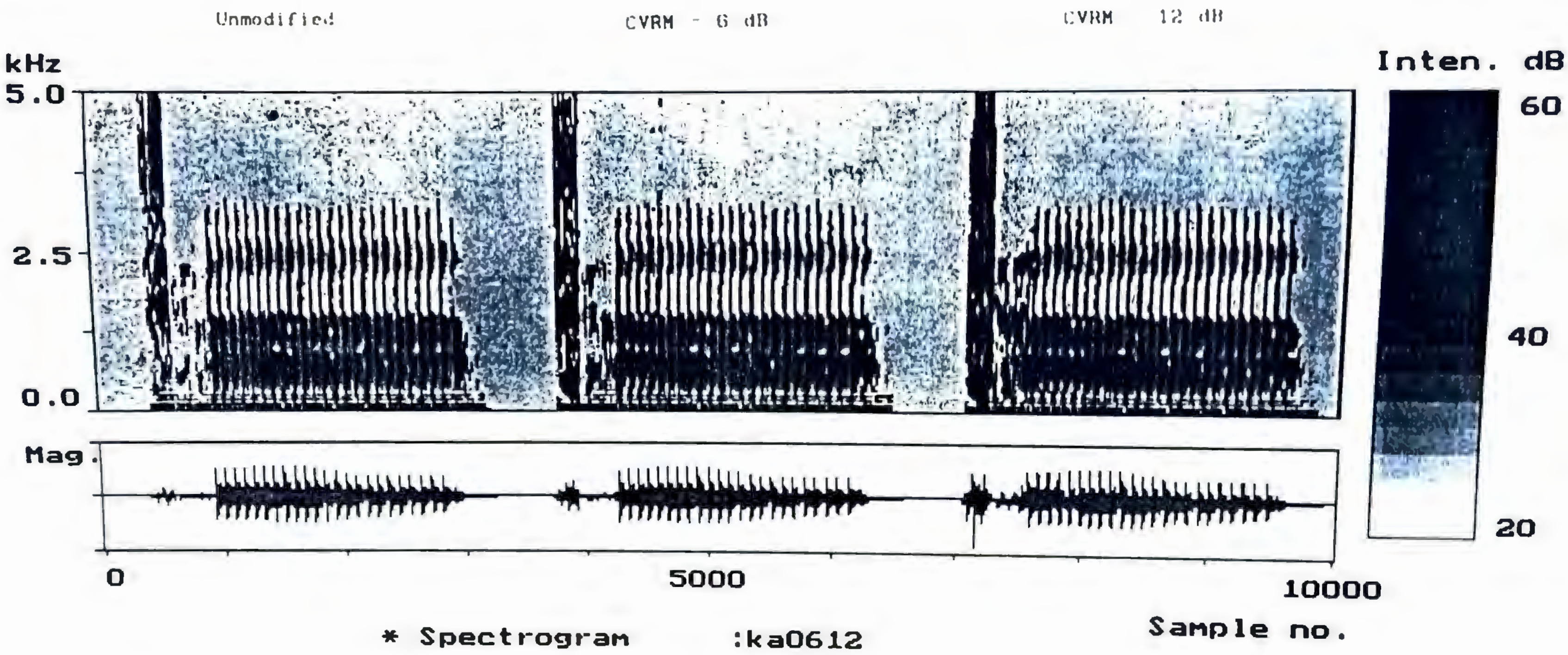


FIG. B.4 Wideband spectrograms for /ka/ for different consonant-to-vowel intensity ratio modifications (CVRMs).

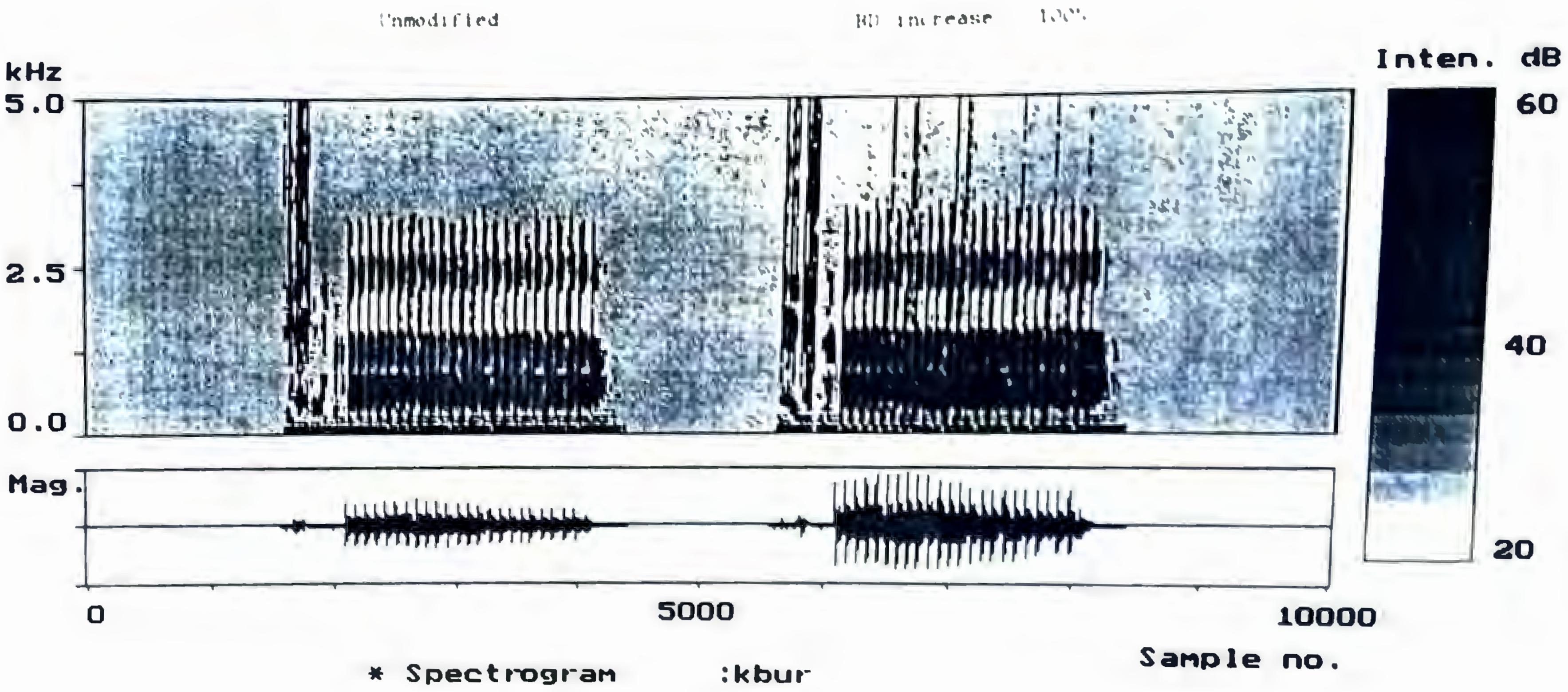


FIG. B.5 Wideband spectrograms for /kɑ/ for different burst duration (BD) values.

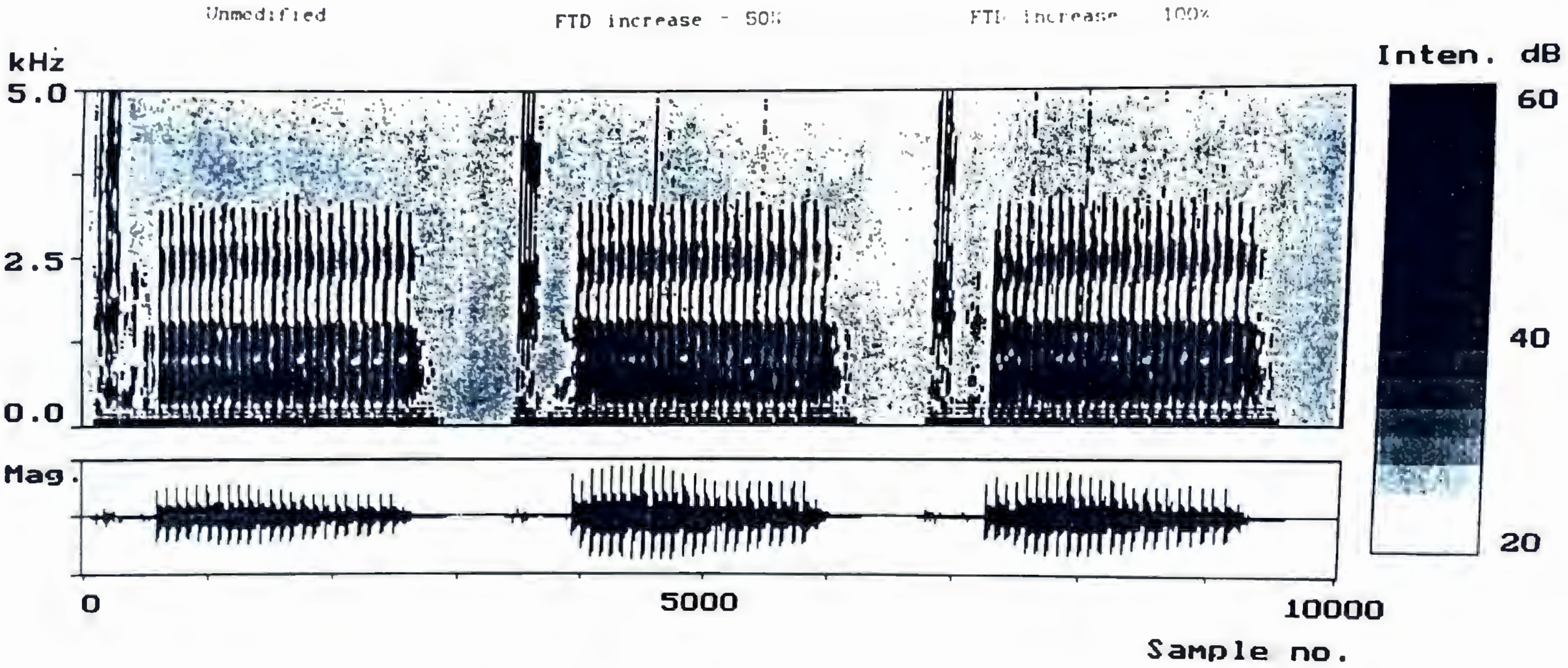


FIG. B.6 Wideband spectrograms for /kɑ/ for different formant transition duration (FTD) values.

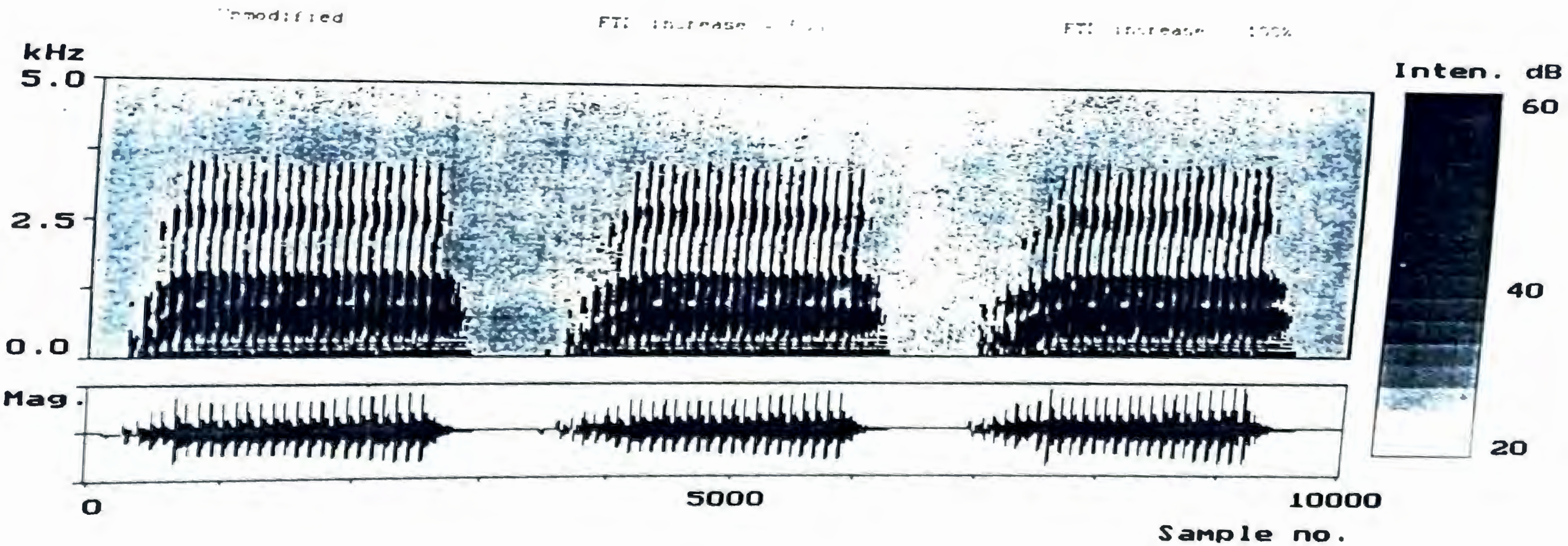


FIG. B.7 Wideband spectrograms for /bː/ for different formant transition duration (FTD) values.

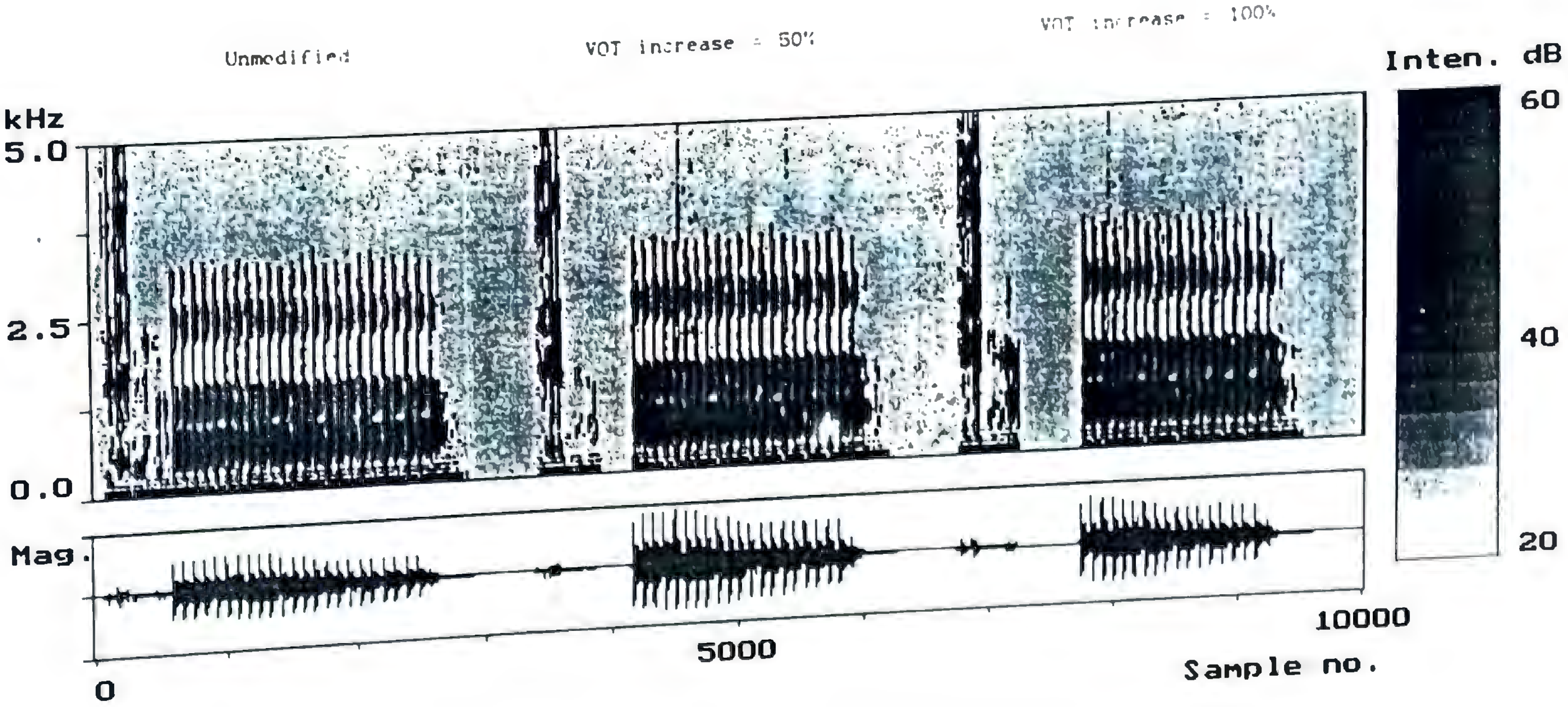


FIG. B.8 Wideband spectrograms for /ka/ for different voice onset time (VOT) values.

APPENDIX C
HARDWARE AND SOFTWARE FOR EXPERIMENT CONTROL

C.1 INTRODUCTION

The aim of the experiments was to evaluate the effect of certain speech processing schemes on the perception of stop consonants by normal-hearing subjects with simulated hearing impairment. The stimuli were to be presented to each subject over a headphone in a randomized order with certain uniformity constraints. Manual administration of the tests has the problem that stimulus items have to be recorded in a number of randomization lists. Hence a computerized test administration system has been developed in order to automate the process. The details of the hardware and software for signal handling and control of experiments are presented below.

C.2 HARDWARE FOR SIGNAL HANDLING

The system was developed for an IBM-PC. Peripherals included a terminal connected to the asynchronous serial port (RS-232) and a PC-based data acquisition card PCL-208 (from Dynalog Micro Systems, Bombay). The terminal was used for displaying the response choices on its screen and for obtaining subject responses from its keyboard.

Some of the important features of the data acquisition card are:

1) Switch selectable 16 single ended or 8 differential A/D

channels with 12-bit resolution

2. Two independent 12-bit D/A channels. Output range of 0-5 V can be created using on-board -5 V reference. External AC or DC reference can also be used to generate other D/A output ranges.

In the present work only one signal channel was needed. A/D channel #0 was used in the -2.5 to +2.5 V range and D/A channel #0 was used in the +5 V range. Assembler subroutines were written for controlling the A/D and D/A operations and these were linked to programs written in C language for carrying out the various signal handling and experiment controlling tasks to be described in the following sections.

For A/D conversion, the input signal should be band-limited to about 5 kHz. Further, a filter is necessary to get a smooth waveform from the staircase waveform obtained at the output of the D/A converter. A seventh-order elliptic low-pass filter [78] was used for this purpose. It has a cutoff frequency of 4.6 kHz. It has a pass band attenuation of 0.3 dB and stop band attenuation of 40 dB. The design and hardware details of this filter are given in Sebastian [99].

The DR-59 headphone was driven by an audio amplifier. The amplifier has facility for driving an headphone, loudspeaker or tape-recorder.

C.3 CONTROL OF EXPERIMENTS

In each experimental run, the stimuli were presented in a randomized order with certain uniformity constraints. For n stimulus items and N presentations in an experiment, these constraints are as follows [78]:

1) Overall uniformity. Each item should be presented N/n times.
2) Short-range uniformity. An item should not be presented more than three times consecutively.
3) Mid-range uniformity. In I consecutive presentations, any item should not occur more than $I/n+2$ times.

For each presentation of the test item, the subject had to respond (guess, if necessary) from the list of choices consisting of all the items in the test set. In order to avoid any bias in responses due to the order in which response choices are provided to the subject, the order of items in the response list and the positions of the correct responses were randomized with uniformity constraints similar to the ones given above.

The FORTRAN program TLIST [78] was used for generating a randomized test list for presenting test stimuli in the experiment. The program reads the input information from a file LIST.DAT in the following format:

Number of test stimuli n ($\leq$ 12),

Number of presentations N ($\leq$ 60),

Two line title for the experiment,

Test item names (each name may be up to six characters, on separate lines).

The program outputs a file SLIST.DAT which will be used by the computerized test administration program, CTA. The stimulus and response items are referred to by order number. The list contains the item to be presented as well as the order in which response choices should be displayed.

A schematic of the experimental set-up was shown in Fig. 5.1. The program for controlling the experiments, CTA, is linked to an assembler routine DA.ASM for controlling the D/A converter. The signal samples for stimuli in the test set are stored in separate files in random binary format. There can be up to 12 items in each experimental run, and the individual items can be upto 20K samples long. The program performs the following tasks:

1) display messages regarding the stimuli and procedure to the the subject,
2) present the stimuli, record the subject response and the response time, display the scores, etc,
3) record the scores as a confusion matrix, display and store the test data for further analysis.

The test information to the program CTA is provided in a file SIGFIL.DAT in the following format:

Sampling frequency in Hz,
Number of items in the test set,
Names of test items (to be displayed to the subject as response choices),
Names of data files containing signal samples for the stimuli.

The information about test presentation is provided in a file SLIST.DAT which is generated by the program TLIST as described earlier.

Before each presentation, choices are displayed on the subject screen. Each response corresponds to a key on the terminal keyboard. The signal data are read from the file, the subject is alerted with a "listen" message on his screen, and presented with the signal over the headphone. The subject has to respond by hitting an appropriate key. There is no timing out, i.e., the subject must respond for the test to proceed. The response is matched with the correct one and is automatically scored. The time taken by the subject in responding is also recorded. The test can be administered with a feedback to the subject by indicating the correct responses or without any such feedback. Usually, the feedback mode is used in the practice sessions and the test data are collected in the no-feedback mode.

As the test progresses, the PC monitor (not the subject screen) gives an update of the keys being hit by the subject, the time taken, cumulative scores, etc. If an invalid key is hit, that particular presentation is deleted from the responses. The test results are stored in a file *sssnn_aa*, where *sss* is the subject identification, *nn* the test number, and aa the attempt number for the experimental condition. The format of the stored results is as under:

Number of test items, number of presentations,

Confusion matrix,

Test number_attempt number, subject and test identification,

Date and time of test,

Number of presentations, number of valid responses, percent score, Minimum, maximum, mean, and standard deviation of response time in sec, total time of the test in minutes.

The confusion matrix results for a test set from several tests for one or more subjects can be combined into a single matrix by using the program CUMMAT.

Appendix D
ANALYSIS OF CONFUSION MATRICES

D.1 INTRODUCTION

A brief description of the programs used for the analysis of confusion matrices is given below. The results of a sample analysis are also presented.

D.2 PROGRAM FOR RECOGNITION SCORES AND INFORMATION TRANSMISSION ANALYSIS

A program INF was written to compute recognition scores and to perform information transmission analysis (as described in Chapter 5) of the input confusion matrix data. The input matrix is usually derived by combining the results from several experimental runs of one or more subjects using the program CUMMAT.

The program reads the input confusion matrix (raw scores, percentage, or fractional relative scores) from the input file in the following format:

1) Test identification in the first two lines.
2) N — no. of stimuli, NT — no. of trials.
3) Stimulus names (0-80 characters). Stimulus names can be one or two characters long, separated by one or more spaces. The first and the last entries are ignored.
4) Confusion matrix cell entries. Entries of a row should be on one line.
5) Minimum, maximum, mean, and standard deviation of the

percentage recognition scores.

6) Minimum, maximum, mean, and standard deviation of response time in sec, total duration of the test in minutes.

The stimulus grouping is read from a file INFOGR.DAT in the following format:

1) N - no. of stimuli, NF — no. of feature groups.

2) Stimulus names (0-80 characters). A stimulus name can be one or two characters long (separated by one or more spaces), and they must be in the same order as in the input confusion matrix.

3) Feature classification information. Feature groups are assigned consecutive integer numbers (-ve, 0, or +ve). Each line should be as follows:

(a) group numbers for stimuli,

(b) name of the feature - upto 15 characters,

(c) group labels (names can be one or two characters long, separated by one or more spaces), upto 20 characters.

The program outputs three files: INFOSC.DAT (recognition scores), INFOTR.DAT (information transmission analysis results), and INFOTS.DAT (summary of both analyses).

D.3 SAMPLE ANALYSIS RESULTS

The inputs to the program INF are the file INFOGR.DAT and the file containing the input confusion matrix data to be analyzed. The stimulus-response confusion matrix selected for this sample analysis corresponds to the experimental condition: CVRM = 9 dB

```
and SNR = 6 dB

File: INFOGR.DAT

6  2
pa ta ka ba da ga
1  2  3  1  2  3   PLACE  LA AL VL
1  1  1  2  2  2   VOICING  UV VO

Input confusion matrix

           OVERALL CONFUSION MATRIX
          (CVRM =  9 dB  SNR =  6 dB)
  6  480
   S\R  pa    ta    ka    ba    da    ga    +
   pa   79     0     0     0     1     0   80
   ta    0    61    15     0     4     0   80
   ka    0     0    80     0     0     0   80
   ba    0     0     0    80     0     0   80
   da    1     3     0     0    69     7   80
   ga    0     0     0     2    17    61   80
    +   80    64    95    82    91    68  480

  81.7  94.4  89.6  5.0
  1.26  2.62  1.95  0.81  4.0
```

The results of the analysis are given below:

```
** RECOGNITION SCORES **

* (6)  OVERALL
S/R  pa    ta    ka    ba    da    ga
pa   99     0     0     0     2     0
ta    0    76    19     0     5     0
ka    0     0   100     0     0     0
ba    0     0     0   100     0     0
da    2     4     0     0    86     9
ga    0     0     0     3    22    76
Correct    89.6%
```

```
• (3): PLACE
S/R   LA    AL    VE
LA    99     1     0
AL     1    86    13
VE     2    10    88
Correct:    91.0%

• (2): VOICING
S/R   UV    VO
UV    97     3
VO     2    98
Correct:   98.1%

** INFORMATION TRANSMISSION **
• (6): OVERALL
Stimulus info. =  2.5841 bits
Response info. =  2.5841 bits
Transn info. =  2.1169 bits
Perc transn. =  81.9

• (3): PLACE
Stimulus info. =  1.5845 bits
Response info. =  1.5845 bits
Transn info. =  1.1612 bits
Perc transn. =  73.3

• (2): VOICING
Stimulus info. =  0.9997 bits
Response info. =  0.9997 bits
Transn info. =  0.8655 bits
Perc. transn. =  86.6
** SUMMARY **
FEATURE    N   COR     IS      IR      IT    RTR
OVERALL    6    90    2.58    2.58    2.12    82
PLACE      3    91    1.58    1.58    1.16    73
VOICING    2    98    1.00    1.00    0.87    87
```

APPENDIX E
EFFECT OF SAMPLE SIZE ON RELATIVE INFORMATION TRANSMITTED

E.1 INTRODUCTION

It has been seen in Chapter 5 that the probabilities of occurrence of the stimulus x_i, response y_j, and stimulus-response pair $(x_i;y_j)$ were estimated from a sample of N observations as given in Eqn. (5.1). The covariance of the stimulus-response, $I(x;y)$, is calculated using these probability values as given in Eqn. 5.5. It has been observed by Miller & Nicely [12] that an estimate based on a finite number of samples will be biased to overestimate $I(x;y)$ for small samples. It was therefore decided to consider the effect of sample size on the results of the information transmission analysis.

E.2 EFFECT OF SAMPLE SIZE

In order to consider the effect of sample size, the stimulus-response confusion matrices obtained for two different experimental conditions in the CV6 test (Chapter 6) have been selected at random and the results are plotted in Figs. E.1 and E.2. For each condition, three of the more stable (in terms of recognition score) confusion matrices were selected for each of the five subjects for analysis. Three sample sizes have been considered here to observe their effect on the overall information transmitted, and information transmitted about place and voicing. In the first case, the fifteen matrices have been considered individually (N = 30), and the scores are plotted in column A of the two figures. In the second case, the three matrices for each subject were combined (N = 90) and the scores are plotted in column B. Column C shows the scores obtained when all the fifteen matrices were combined to yield a single matrix (N = 450). The average score (and its standard deviation) is also indicated for all the conditions. It may be noted that some of the points represent coincident scores obtained by different subjects.

It is observed from the figures that there is indeed an overestimate in the scores obtained for lower values of *N* and that the scores decrease as *N* is increased. Thus it is seen that an estimate based on a finite number of samples will be biased to overestimate the scores for small sample size. The spread in scores is seen to be higher for place and voicing information than for overall information transmitted.

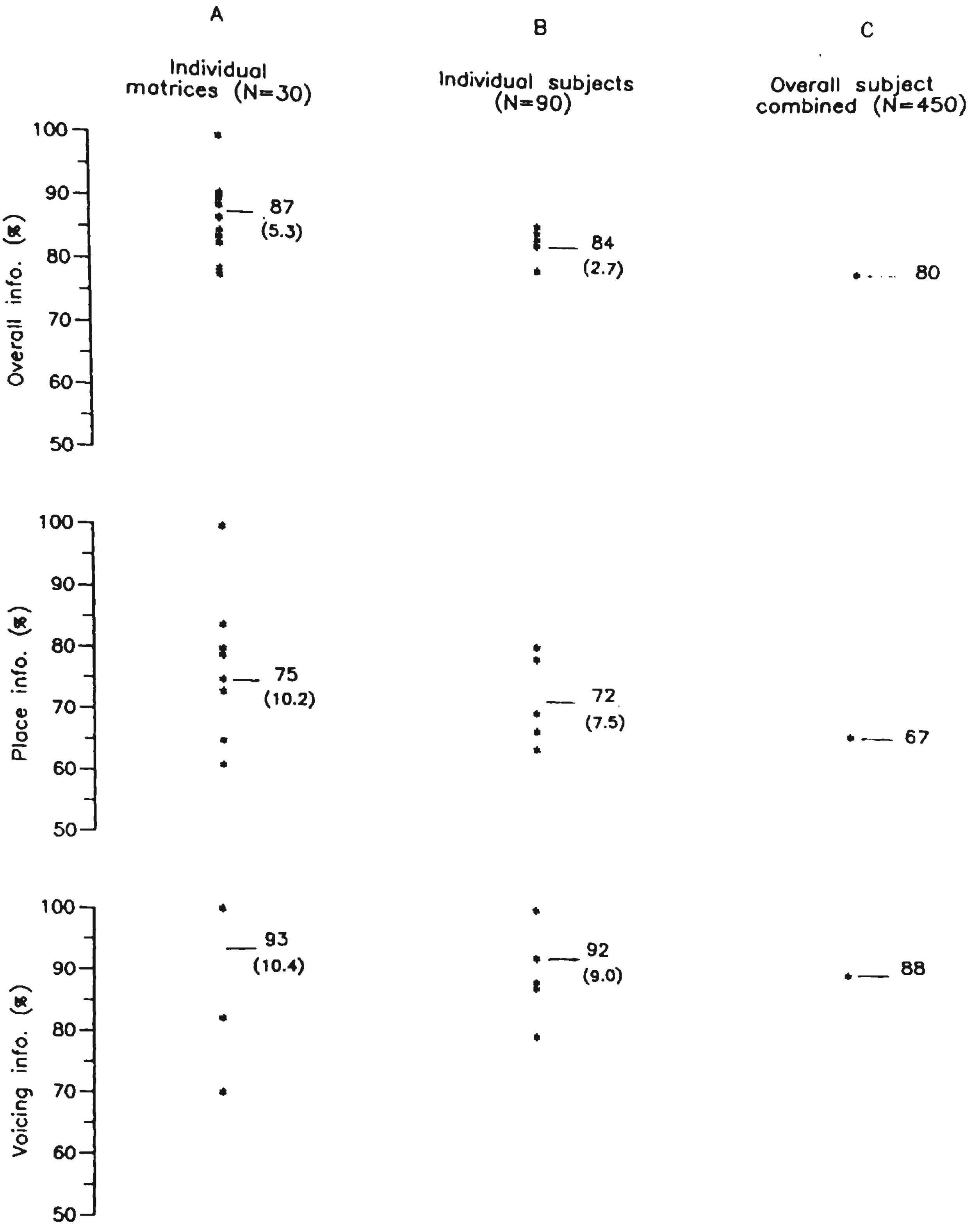


FIG. E.1 Effect of sample size, N, on the results of information transmission analysis. (Test CV6 – CVRM : 6 dB; SNR : 12 dB).

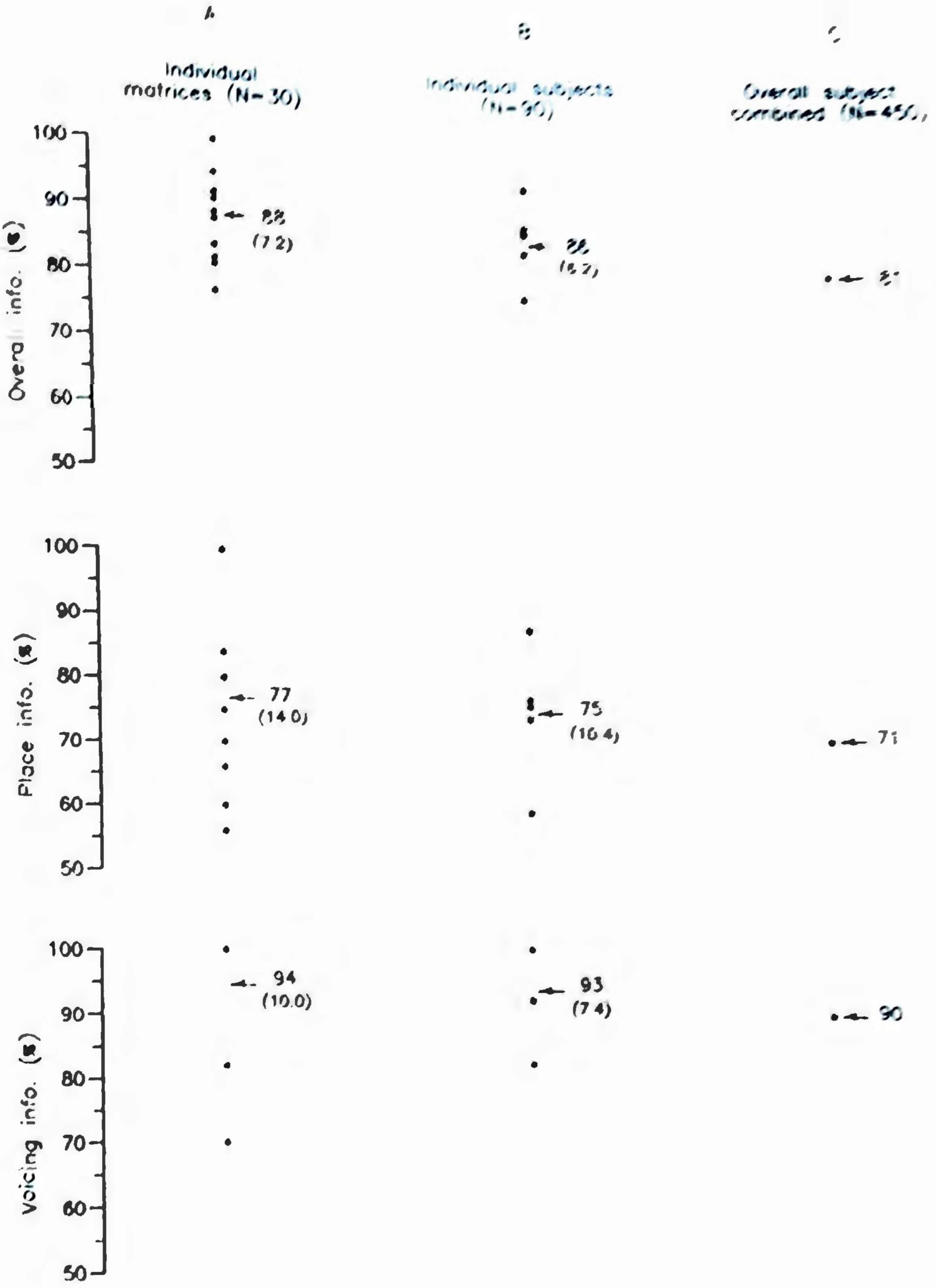


FIG. E.2 Effect of sample size, N, on the results of information transmission analysis (Test CV6 – CVRM : 12 dB; SNR : 6 dB).

APPENDIX F
CALIBRATION OF DR-59 HEADPHONE

F.1 INTRODUCTION

The headphone most commonly used as standard in the area of speech perception experiments and in audiological testing belongs to the TDH-series; e.g., TDH-39 or TDH-49. However, these headphones are difficult to procure in the Indian market. A majority of the Indian audiometers are supplied with the DR-59 Elega headphone (Japanese make) with rubber cushions. A DR-59 headphone was procured for use in the experiments reported here.

It was important that the test stimuli be presented at a calibrated sound level. This would have been possible if the stimuli were presented through an audiometer. Due to the non-availability of an audiometer in our lab it was decided to perform electroacoustic calibration of the DR-59 headphone using the facilities available at the Ali Yavar Jung National Institute for the Hearing Handicapped at Bandra, Bombay. The electroacoustic properties of the DR-59 headphone were determined (as described in the next section) in order to select the earpiece which matched the TDH-39 characteristics more closely.

F.2 CALIBRATION PROCEDURE

F.2.1 Equipment used

1) Audiometer: Qualitone Acoustic Appraiser, ANSI-1969.
2) Sound Level Meter: Type 2230 B&K Precision Integrating Sound Level Meter, along with Type 1625 B&K 1/3-1/1 Octave Filter

Set, and Type 4155 B&K 1/2-inch Condenser Microphone This entire system was calibrated with a Type 4230 B&K Sound Level Calibrator (94 SPL at 1 kHz).

3) 6 cc Coupler in Artificial Ear.

4) HP Multimeter Type 34401A to obtain the equivalent voltage across the earpiece input for a given sound level at the earpiece output.

5) Junction Box to tap the earpiece input.

F.2.2 Method

The set-up used for calibrating the audiometer and obtaining the electroacoustic characteristics of the headphone is shown in Fig. F.1. The earpiece is placed over a carefully machined coupler, usually containing a cavity of precisely 6 cm^3. A weight of 500 grams or a spring with equivalent tension holds the receiver (earpiece) in place. Sounds emanating from the diaphragm of the receiver are picked up by a sensitive microphone at the bottom of the coupler (often called the artificial ear), amplified, and read in dB SPL on a sound-level meter. In order to be certain that the meter is reading the level of the tone (or other signal) from the receiver and not from the ambient room noise, the level of the signal is usually high enough to avoid this possibility [100]. Hearing levels of 70 dB are convenient for this purpose, and the readout should correspond to the number of decibels required for threshold of the particular signal plus 70 dB.

Since the speech perception tests were to be conducted in the absence of an audiometer, the voltage level at the earpiece input corresponding to each reading was also noted. The voltage was measured by tapping the stereo audio jack from the audiometer using a junction box made for the purpose. The pure tone electroacoustic characteristics for the TDH-39 and the pair of DR-59 earpieces were obtained and a combined plot of the earpiece input voltage, in dBm, versus frequency for a 100 dB SPL at the earpiece output is shown in Fig. F.2. (The dBm operation calculates the power delivered to a 600 ohm resistance referenced to 1 mw). It is observed from the plot that the DR-59b earpiece matched the characteristic of the TDH-39 earpiece more closely than the DR-59a earpiece. Hence the DR-59b earpiece was selected for the experiments reported in this thesis.

Since the test stimuli were speech sounds it was necessary to calibrate the audiometer in the speech mode. For this, a pure tone, a noise containing approximately equal intensities at all frequencies, or a sustained vowel sound, may be used as the input. Here, a synthesized vowel /a/ was presented repetitively in order to simulate a sustained vowel. This signal was adjusted in amplitude so that the VU meter on the audiometer read zero. With the hearing-level dial set at 70 dB and with the earpiece on the coupler, the signal should read 90 dB SPL on the meter (70 dB HL plus 20 dB SPL required for audiometric zero for speech on the ANSI-1969 standard) for the TDH-39 earpiece. For this calibrated speech audiometer, the earpiece input voltage versus ouput SPL readings were obtained using the DR-59b earpiece and a best-fit plot was obtained as shown in Fig. F.3.

In the present study, all the stimuli were presented at a level of about 75 dB SPL. As seen from Fig. F.3, 75 dB SPL corresponds to an earpiece input voltage of -36 dBm. Hence, at the start of each session the volume control on the power amplifier section driving the headphone was adjusted (if necessary) so that the earpiece input voltage for a repetitive presentation of the synthesized vowel /a/ was 0 dBm.

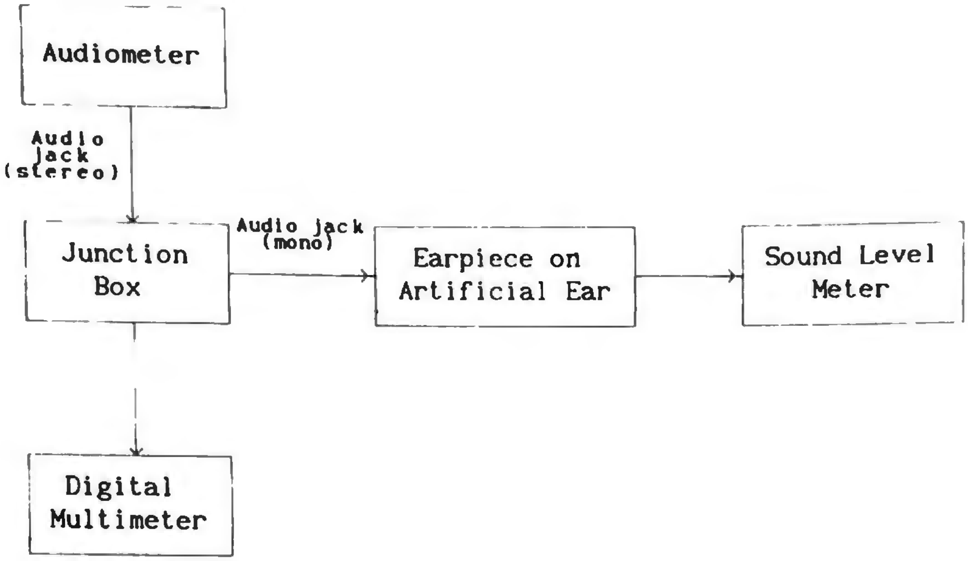


FIG. F.1 Set-up for electroacoustic calibration of headphone.

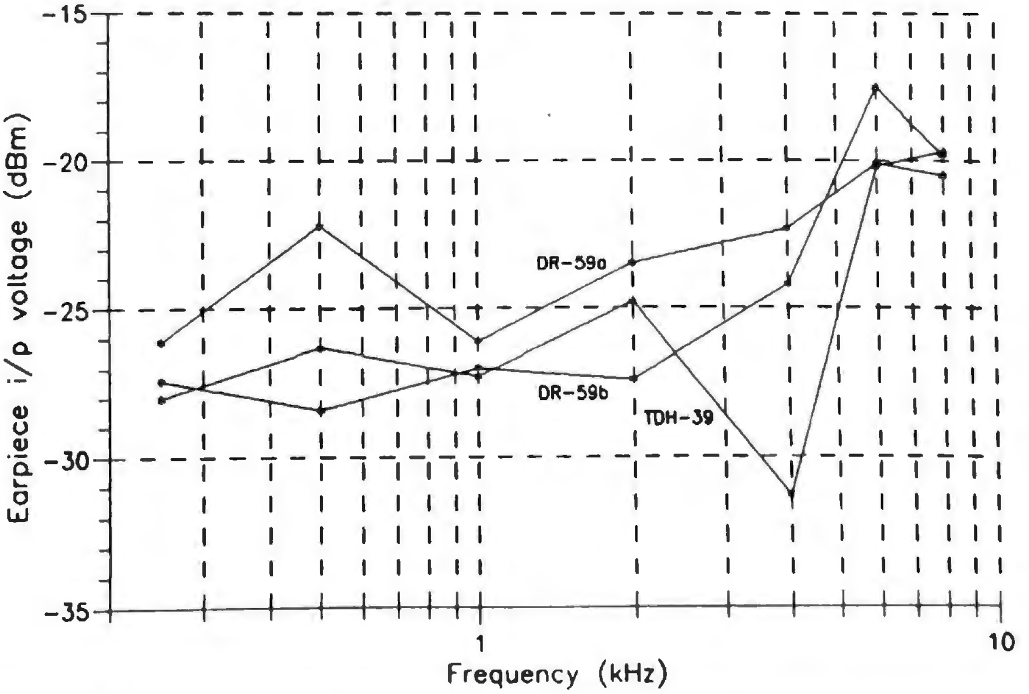


FIG. F.2 Earpiece input voltage versus frequency for 100 dB SPL at earpiece output.

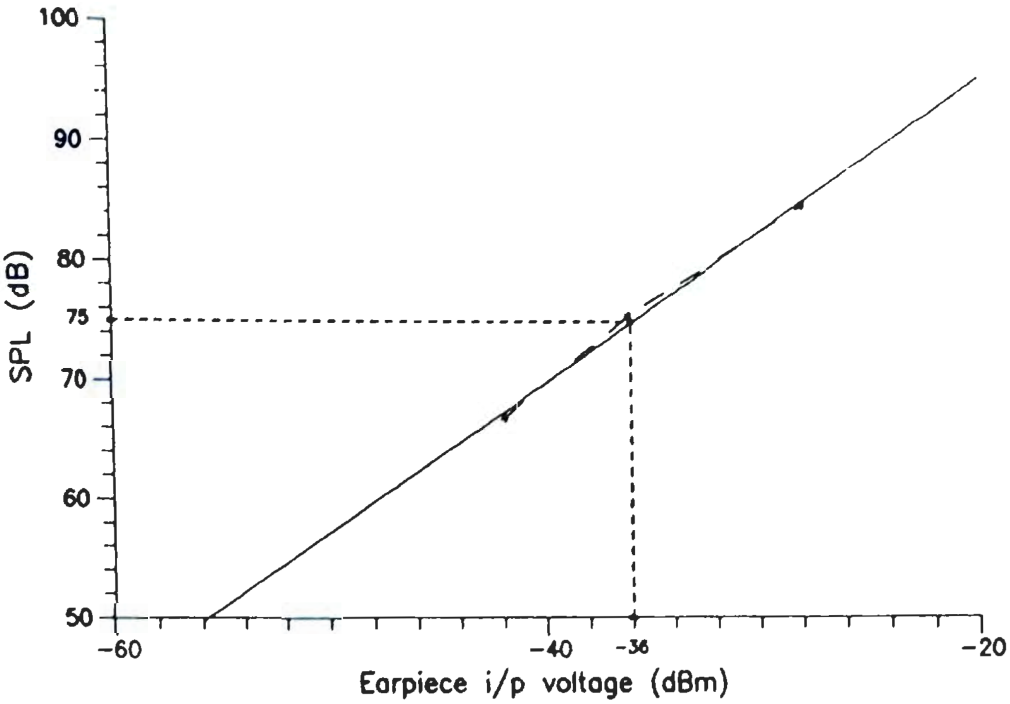


FIG. F.3 Output SPL versus input voltage for DR-59 earpiece.

APPENDIX G
SUBJECT DATA

G.1 SCALE OF HEARING IMPAIRMENT AND SUBJECT DATA

An approach for describing impaired hearing utilizes the monoral pure tone threshold average (PTA) in the speech frequencies and attaches subjective descriptors to the resultant levels. These suggestions are summarized in Table G.1.

The five subjects chosen for the present investigations underwent an hearing screening test at the Ali Yavar Jung National Institute for the Hearing Handicapped, Bombay, and the results of the test are given in Table G.2. It is observed that all the subjects have pure tone auditory thresholds within 20 dB of the normal hearing standards.

Table G.1 Scale of hearing impairment. Source: [101]

Average threshold level (dB) (*re* ANSI-1969 for 0.5, 1 and 2 kHz).	Suggested description
-10 to 25	Normal hearing
26 to 40	Mild hearing loss
41 to 55	Moderate hearing loss
56 to 70	Moderately severe hearing loss
71 to 90	Severe hearing loss
91 plus	Profound hearing loss

TABLE 6.2 Results of hearing screening test performed on the five subjects. PTA: Pure tone average hearing threshold level (test frequencies 0.5, 1, 2 kHz) in dB HL

Subject Code	Ear L=left R=right	Hearing Threshold (dB) Frequency (kHz) 0.25	0.5	1	2	4	6	8	PTA (dB)
S1	L	10	15	20	15	10	25	20	17
	R	15	15	10	10	15	10	10	12
S2	L	20	20	20	20	20	5	0	20
	R	20	20	20	20	20	15	5	20
S3	L	20	25	20	20	20	15	15	22
	R	20	20	10	10	5	20	15	13
S4	L	15	10	10	25	15	15	20	15
	R	40	20	10	20	5	25	20	20
S5	L	5	10	15	10	15	15	10	12
	R	10	15	10	10	10	20	10	12

Appendix H
TEST INSTRUCTIONS

H.1 Test instructions to the normal hearing subjects.

The purpose of these experiments is to evaluate certain speech processing schemes on the perception of stop consonants.

Your task will be to listen to and identify the syllable presented. The sounds will be presented monaurally (in one ear only) over the headphone. You will be seated in front of a computer terminal and will use the keyboard to indicate your response after each stimulus presentation. The number of test sounds may vary from 6 to 9, and a single test will take typically 5 - 8 minutes. A test session will involve several tests and may take 2 - 3 hours; however, you may request for a break at an earlier time.

Instructions for a particular test will be displayed on your terminal screen at the start of the test session. In the beginning you may undergo a trial run with correct answer feedback in order to become familiar with the set. You should listen to these sounds several times in order to establish an association between the sounds presented and the names used to identify them.

During the test, the presentation number and the set of choices will be displayed (in a random order). A "listen" message will be displayed before each presentation. You will indicate your response by hitting the appropriate key on the keyboard. The next presentation will follow after a brief pause (2-5 seconds). A presentation will not be repeated. If you are not sure, you can guess. The test will not proceed if you do not respond. If you missed a presentation, you may indicate this by hitting a key other than the valid choices.

H.2 Test instructions as displayed on the subject terminal screen during the CV6 test.

**** CONSONANT IDENTIFICATION TEST ****

Your task is to identify the presented sound from among the following:

/pa/, /ta/, /ka/, /ba/, /da/, /ga/

After listening to the sound, please hit the corresponding key as quickly as possible

A presentation will not be repeated. If you are not sure, you can guess. The test will not proceed if you do not respond. If you missed a presentation, and cannot even guess, you may hit a key other than the valid choices.

PLEASE HIT ANY KEY WHEN YOU ARE READY FOR THE TEST

REFERENCES

[1] F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.*, vol. 24, pp. 597-606, 1952

[2] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 64, pp. 1358-1368, 1978.

[3] J. R. Dubno, D. D. Dirks, and L. R. Langhofer, "Evaluation of hearing-impaired using the nonsense syllable test: II," *J. Speech Hear. Res.*, vol. 25, pp. 141-148, 1982.

[4] M. C. Narasimhan and A. K. Mukherjee, *Disability - A Continuing Challenge*. New Delhi: Wiley Eastern Limited, 1986.

[5] CHABA, "Speech-perception aids for the hearing-impaired people: Current status and needed research," *J. Acoust. Soc. Am.*, vol. 90, pp. 637-685, 1991.

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. London: Academic Press, 1982.

[7] M. F. Dorman, K. Marton, M. T. Hannley, and J. M. Lindholm, "Phonetic identification by elderly normal and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 77, pp. 664-670, 1985.

[8] B. Johansonn, "The use of the transposer for the management of the deaf child," *Internat. Audiol.*, vol. 5, pp. 362-373, 1966.

[9] M. A. Picheny, *Speaking Clearly for the Hard of Hearing.* Ph. D. Thesis, MIT, Cambridge, Massachusetts, 1981.

[10] H. Ono, T. Okasaki, S. Nakai, and H. Harasaki, "Identification of an emphasized consonant of a monosyllable in hearing-impaired and its application to a hearing aid," *J. Acoust. Soc. Am.*, vol. 71, S58, 1982.

[11] S. Gordon-Salant, "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *J. Acoust. Soc. Am.*, vol. 80, pp. 1599-1607, 1986.

[12] G. A. Miller and P. E. Nicely, " An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338-352, 1955.

[13] H. Levitt, J. M. Pickett, and R. A. Houde, Eds, *Sensory Aids for the Hearing Impaired.* New York: IEEE Press, 1980.

[14] J. O. Pickles, *An Introduction to the Physiology of Hearing.* London: Academic, 1982.

[15] E. M. Danaher, M. Wilson, and J. M. Pickett, "Backward and forward masking in listeners with severe sensorineural hearing loss," *Audiology*, vol. 17, pp. 324-338, 1978.

[16] H. K. Dunn and S. D. White, "Statistical measurements in

conversational speech," *J. Acoust. Soc. Am.*, vol. 11, pp. 278-288, 1940.

[17] P. J. Fitzgibbons and F. L. Wightman, "Gap detection in normal and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 72, pp. 761-765, 1982.

[18] R. S. Tyler, Q. Summerfield, E. J. Wood, and M. A. Fernandes, "Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 72 pp. 740-752, 1982.

[19] S. Revoile, L. Holden-Pitt, J. Pickett, and F. Brandt, "Speech cue enhancement for the hearing impaired: I. Altered vowel durations for perception of final fricative voicing," *J. Speech Hear. Res.*, vol. 29, pp. 240-255, 1986.

[20] N. Hannley and M. F. Dorman, "Susceptibility to intraspeech masking in listeners with sensorineural hearing loss," *J. Acoust. Soc. Am.*, vol. 74, pp. 40-51, 1983.

[21] M. Leek, M. Dorman, and Q. Summerfield, "Minimal spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 81, pp. 148-154, 1987.

[22] I. B. Crandall, "The composition of speech," *Phys. Rev.*, vol. 10(2), 1917. Reprinted in M. E. Hawley, Ed, *Speech Intelligibility and Speaker Recognition*. Stroudsberg, PA:

Dowden Hutchinston Ross, 1977

[23] J. J. Godfrey and K. K. Millay, "Perception of synthetic speech sounds by hearing-impaired listeners," *J. Aud. Res.*, vol. 20, pp. 187-203, 1980.

[24] M. F. Dorman and K. Marton, "Cochlear frequency selectivity and phonetic identification in aging listeners," *J. Acoust. Soc. Am.*, vol. 69, S123, 1981.

[25] E. Owens, M. Benedict, and E. D. Schubert, " Consonant phonemic errors associated with pure tone configurations and certain kinds of hearing impairment," *J. Speech Hear. Res.*, vol. 15, pp. 308-322, 1972.

[26] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Am.*, vol. 54, pp. 1248-1266, 1973.

[27] L. C. W. Pols and M. E. H. Schouten, "Identification of deleted consonants," *J. Acoust. Soc. Am.*, vol. 64, pp. 1333-1337, 1978.

[28] J. R. Dubno and H. Levitt, "Predicting consonant confusions from acoustic analysis," *J. Acoust. Soc. Am.*, vol. 69, pp. 249-261, 1981.

[29] S. A. Gelfand, N. Piper, and S. Silman, "Consonant recognition in quiet as a function of aging among normal hearing subjects," *J. Acoust. Soc. Am.*, vol 78,

pp. 1198-1206, 1986.

[30] W. J. Staab and S. F. Lybarger, "Characteristics and use of hearing aids," p. 657 in J.Katz, Ed, *Handbook of Clinical Audiology*. Baltimore: Williams and Wilkins, 1994

[31] E. Douek, A. J. Fourcin, B. C. J. Moore, and G. P. Clarke, "A new approach to the cochlear implant," *Proc. Roy. Soc. Med.*, vol. 70, pp. 379-383, 1977. Reprinted in H. Levitt, J. M. Pickett, and R. A. Houde, Eds, *Sensory Aids for the Hearing Impaired.*, pp. 460-464, New York: IEEE Press, 1980.

[32] J. H. Kirman, " Tactile communication of speech: a review and an analysis," *Psychol. Bull.*, pp. 54-74, 1974. Reprinted in H. Levitt, J. M. Pickett, and R. A. Houde, Eds, *Sensory Aids for the Hearing Impaired.*, pp 297-317, New York: IEEE Press, 1980.

[33] J. Montano, "Rehabilitation technology for the hearing impaired," p. 638 in J. Katz, Ed, *Handbook of Clinical Audiology*. Baltimore: Williams and Wilkins, 1994.

[34] L. D. Braida , "Speech processing for the hearing impaired," *Proc. Second Int. Conf. Rehab. Engg.*, Ottawa, 1984, pp. 146-149.

[35] E. Villchur, "Signal processing to improve speech intelligibility in perceptive deafness," *J. Acoust. Soc. Am.*, vol. 53, pp. 1646-1657, 1973.

[36] R. Lippman, L. Braida, and N. I. Durlach, "Study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss," *J. Acoust. Soc Am.*, vol. 69, pp. 524-534, 1981.

[37] I. V. Nabelek, "Performance of hearing-impaired listeners under various types of amplitude compression," *J. Acoust. Soc. Am.*, vol. 74, pp. 776-791, 1983.

[38] P. L. Branderbit, "A standardized programming system and three-channel compression hearing instrument technology," *Hear. Instrum.*, vol. 42, pp. 24-30, 1991.

[39] R. Abramovitz, "Frequency shaping and multiband compression in hearing aids," *J. Acoust. Soc. Am.*, vol. 65, S136, 1979.

[40] D. R. Allen, W. T. Strong, and E. P. Palmer EP, "Experiments on the intelligibility of low frequency codes," *J. Acoust. Soc. Am.*, vol. 70, pp. 1248-1255, 1981.

[41] D. Ling(1969). Speech discrimination by profoundly deaf children using linear and coding amplifiers," *IEEE Trans. Aud. Electroacoust.*, vol. 17, pp. 298-303, 1969.

[42] K. O. Foust and R. W. Gengel, "Speech discrimination by sensorineural hearing-impaired persons using a transposer hearing aid," *Scand. Audiol.*, vol. 2, pp 161-170, 1973

[43] C. M. Reed, B. L. Hicks, L. D Braida, and N I Durlach,

"Discrimination of speech processed by low-passed filtering and pitch-invariant frequency lowering," *J. Acoust. Soc. Am.*, vol 74, pp. 409-419, 1983.

[44] A S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, " Articulation testing methods: Consonantal differentiation with a closed response set," *J. Acoust. Soc. Am.*, vol. 37, pp. 158-166, 1965.

[45] R. L. Freyman and G. P. Nerbonne, "The importance of consonant-vowel intensity ratio in the intelligibility of voiceless consonants," *J. Speech Hear. Res.*, vol. 32, pp. 524-535, 1989.

[46] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *J. Speech Hear. Res.*, vol. 28, pp. 96-103, 1985.

[47] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.*, vol. 29, pp. 434-446, 1986.

[48] A. A. Montgomery and R. A. Edge, "Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults," *J. Speech Hear. Res.*, vol. 31, pp. 386-393, 1988.

[49] S. Gordon-Salant, "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects *J. Acoust. Soc. Am.*, vol. 81, pp. 1199-1202, 1987.

[50] S. Revoile, L. Holden-Pitt, D. Edward, J. Pickett, and F. Brandt, "Speech-cue enhancement for the hearing-impaired: II. Amplification of burst/murmur cues for improved perception of final stop voicing," *J. Rehab. Res. Dev.*, vol. 24, pp. 207-216, 1987.

[51] R. L. Freyman, G. P. Nerbonne, and H. A. Cote, "Effect of consonant-vowel ratio modification on amplitude envelope cues for consonant recognition," *J. Speech Hear. Res.*, vol. 34, pp. 415-426, 1991.

[52] G. Fairbanks, W. L. Everett, and R. P. Jaeger, "Method for time or frequency compression-expansion of speech," *IRE Trans. Audio*, vol. AU-2, p. 7, 1954.

[53] T. D. Schon, "Effects of speech intelligibility of time-compression and expansion on normal hearing, hard of hearing, and aged males," *J. Aud. Res.*, vol. 10, p. 263, 1970.

[54] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *J. Acoust. Soc. Am.*, vol. 55, pp. 653-659, 1974.

[55] D. J. Van Taseil, L. T. Hagen, L. L. Koblas, and

S. G. Penner, "Perception of short-term spectral cues for stop consonant place by normal and hearing-impaired subjects," *J. Acoust. Soc. Am.*, vol. 72, pp. 1771-1780, 1982.

[56] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971-995, 1980.

[57] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. New York: Springer-Verlag, 1972.

[58] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[59] R. Koenig, H. K. Dunn, and L. Y. Lacey, "The sound spectrograph," *J. Acoust. Soc. Am.*, vol. 18, pp. 19-49, 1946.

[60] T. G. Thomas, P. C. Pandey, and S. D. Agashe, " A PC-based spectrograph for speech and biomedical signals," *Proc. Int. Conf. Recent Advances Biomed. Engg.*, Hyderabad, 1994, pp. 7-10.

[61] T. G. Thomas, P. C. Pandey, and S. D. Agashe, "A PC-based multiresolution spectrograph," *J. IETE*, vol. 40, pp. 105-108.

[62] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Massachusetts: Addison-Wesley, 1987.

[63] M. Halle, G. W. Hughes, and J-P A Radley, "Acoustic properties of stop consonants," *J. Acoust. Soc. Am.*, vol. 29

pp 107-116, 1957.

[64] P. Ladefoged, *A Course in Phonetics*. New York: Harcourt Brace Jovanivich, 1982.

[65] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Rev.*, vol. 74, pp. 431-464, 1967.

[66] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Am.*, vol. 67, pp. 648-662, 1980.

[67] D. Kewley-Port, "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," *J. Acoust. Soc. Am.*, vol. 72, pp. 379-389, 1982.

[68] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 769-773, 1955.

[69] K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, " Effect of third formant transitions on the perception of voiced stop consonants," *J. Acoust. Soc. Am.*, vol. 30, pp. 122-126, 1958.

[70] D. B. Fry, *The Physics of Speech*. Cambridge: Cambridge University Press, 1979.

[71] V. W. Zue, *Acoustic Characteristics of Stop Consonants: A*

Controlled Study. D Sc. thesis, MIT, Cambridge Massachusetts, 1976.

[72] B. H. Repp, "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," *Lang. Speech*, vol. 27, pp. 173-189, 1979.

[73] S. A. Chafekar, *Speech Synthesis for the Testing of Sensory Aids for the Hearing-Impaired*. M. Tech. thesis, School of Biomedical Engineering, IIT, Bombay, 1990.

[74] D. M. Kulkarni, *Development of a Cascade Pole-Zero Speech Synthesizer*. M. Tech. thesis, Department of Electrical Engineering, IIT, Bombay, 1992.

[75] N. N. Raut, *A Text to Speech Converter Using Cascaded Pole Zero Synthesizer*. M. Tech. thesis, Department of Electrical Engineering, IIT, Bombay, 1994.

[76] J. M. Pickett, *The Sounds of Speech Communication*. Baltimore: University Park, 1980.

[77] P. E. Papamichalis, *Practical Approaches to Speech Coding*. New Jersey: Prentice Hall, 1987.

[78] P. C. Pandey, *Speech Processing for Cochlear Prostheses*. Ph. D. thesis, Department of Electrical Engineering, University of Toronto, 1987.

[79] C. Simon, "On the use of comfortable listening levels in speech experiments," *J. Acoust. Soc. Am.*, vol. 64,

pp. 744-751, 1978

[80] M. F. Dorman and K. Dougherty, "Shifts in phonetic identification with changes in signal presentation level," *J. Acoust. Soc. Am.*, vol. 69, pp. 1439-1440, 1981.

[81] D. J. Van Tasell and E. B. A. Crump, "Effects of stimulus level on perception of two acoustic cues of speech," *J. Acoust. Sc. Am.*, vol. 70, pp. 1527-1529, 1981.

[82] L. E. Humes, D. D. Dirks, T. S. Bell, and G. E. Kincaid, "Recognition of nonsense syllables by hearing-impaired listeners and by noise-masked normal listeners," *J. Acoust. Soc. Am.*, vol. 81, pp. 765-773, 1987.

[83] S. DeGennaro, L. D. Braida, and N. I. Durlach, "Study of multi-band syllabic compression with simulated sensorineural hearing loss," *J. Acoust. Soc. Am.*, vol. 69, S16, 1981.

[84] H. Fletcher, "The perception of sounds by deafened persons," *J. Acoust. Soc. Am.*, vol. 24, pp. 490-497, 1952.

[85] J. P. A. Lochner and J. F. Burger, "Form of the loudness function in the presence of masking noise," *J. Acoust. Soc. Am.*, vol. 33, pp 1705-1707, 1961.

[86] L. E. Humes, B. Espinoza-Varas, and C. S. Watson, "Modeling sensorineural hearing loss. I. Model and retrospective evaluation," *J. Acoust. Soc. Am.*, vol. 83, pp. 188-202, 1988.

[87] M. Florentine and S. Buus, "Temporal gap-detection in

sensorineural and simulated hearing impairments," *J. Speech Hear. Res.*, vol. 27, pp. 449-455, 1984.

[88] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379-423, 1948.

[89] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10-21, 1949.

[90] R. Ahmed and S. S. Agrawal, "Significant features in the perception of (Hindi) consonants," *J. Acoust. Soc. Am.*, vol. 45, pp. 758-763, 1969.

[91] J. P. Gupta and A. Ahmad, "Perception of (Hindi) consonants in differentiated clipped speech," *J. Acoust. Soc. Am.*, vol. 58, pp. 282-283, 1975.

[92] R. C. Bilger and M. D. Wang, "Consonant confusions in patients with sensorineural hearing loss," *J. Speech Hear. Res.*, vol. 19, pp. 718-748, 1976.

[93] C. W. Turner and M. P. Robb, "Audibility and recognition of stop consonants in normal and hearing-impaired subjects," *J. Acoust. Soc. Am.*, vol. 81, pp. 1566-1573, 1987.

[94] B. H. Repp, "On the levels of description in speech research," *J. Acoust. Soc. Am.*, vol. 69, pp. 1462-1464, 1981.

[95] T. G. Thomas, P. C. Pandey, and S. D. Agashe, "On the importance of consonant-vowel intensity ratio in speech enhancement for the hearing impaired," *Proc. Int. Conf.*

Biomed. Engg., Hong Kong, 1994, pp. 181-184.

[96] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, pp. 51-66, 1978.

[97] L. Cohen, "Time-frequency distributions — a review," *Proc. IEEE*, vol. 77, pp. 941-981, 1989.

[98] S. Cheung and J. S. Lim, "Combined multiresolution (wideband/ narrowband) spectrogram," *IEEE Trans. Sig. Proc.*, vol. 40, pp. 975-977, 1992.

[99] G. Sebastian, *A Speech Training Aid for the Hearing Impaired*. B. Tech. Project Report, Department of Electrical Engineering, IIT, Bombay, 1991.

[100] F. N. Martin, *An Introduction to Audiology*. New Jersey: Prentice-Hall, 1975.

[101] P. A. Yantis, "Pure tone air-conduction testing," p. 164 in J. Katz, Ed, *Handbook of Clinical Audiology*. Baltimore: Williams and Wilkins, 1994.

[102] B. N. Ashok Baragi, *A Speech Training Aid for the Deaf*. M. Tech. Thesis, Department of Electrical Engineering, IIT, Bombay, 1996.

[103] V. V. Siva Rama Prasad, *Speech Processing for Single Channel Sensory Aid*. M. Tech. Thesis, Department of Electrical Engineering, IIT, Bombay, 1996.

SUMMARY

A number of research studies have turned to speech enhancement as a possible means of increasing intelligibility for persons with moderate to severe sensorineural hearing loss. One promising scheme for enhancing the speech signal was based on studies of speaking clearly for the hearing impaired. Acoustic analysis revealed that there were large differences in the temporal characteristics of clear speech compared to conversational speech. Two specific characteristics of clear speech that were selected for evaluation in the present investigation were consonant-to-vowel (C/V) intensity ratio and consonantal duration. These two characteristics are capable of being manipulated independently in the time domain.

A few studies reported in the literature have evaluated the intelligibility of natural speech artificially transformed to clear speech by altering either one or both of the above acoustic parameters. While all the studies reported improvement in consonant recognition for C/V ratio modification, albeit by different amounts, the effect of increasing consonantal duration has been equivocal. This may have been because the acoustic segments associated with consonant phonemes increase in a nonuniform manner and such changes cannot be aptly simulated by transforming natural speech to clear speech.

In the present investigation, we have used synthetic test stimuli generated using a modified version of the Klatt synthesizer. Since the synthesis parameters were under user control, the segmentation of individual phonemes is more accurate and it is possible to alter the various acoustic segments independently. In order to keep the test set as well as the feature contrasts simple, the consonant phonemes were limited to the English stop consonants /p, t, k, b, d, g/.

The aim of the present investigation was to study the extent of consonant-to-vowel intensity ratio (CVR) modification and

consonantal duration (CD) modification that would be helpful in speech perception by the hearing impaired, listening tests were conducted on five normal-hearing persons both in the quiet and under simulated hearing-impairment conditions. Hearing impairment was simulated by mixing the stimuli with broadband masking noise. Studies reported in the literature have shown that in addition to simulating reduced dynamic range, broadband noise closely approximates the loudness growth function of listeners with sensorineural hearing loss. The tests were performed under three listening conditions: no masking noise, and masking noise with 12 dB and 6 dB SNR.

Due to the repetitive nature of the experiments, a computerized test administration system was developed in order to automate the process. The subject was seated in front of a terminal which was connected to the asynchronous serial port of a PC which controlled the experiment. The terminal was used for displaying the response choices and for obtaining the subject response from its keyboard. The acoustic stimuli were presented through a D/A converter, amplifier, and an audiometric headphone to the subject. Each experimental run consisted of a number of presentations of the stimuli, in a randomized order with certain uniformity constraints. In order to minimize subject response bias, the order of items in the response list as well as the positions of the correct responses were randomized. At the end of each run, the stimulus-response confusion matrix, the recognition score, and the mean response time were stored.

For evaluation purposes, the stimulus-response confusion matrices from a number of test runs for each experimental condition were combined, and two different measures based on them were used: (a) recognition score and (b) relative information transmission. While recognition scores are easiest to calculate and interpret, information transmission analysis has the merit that it measures the covariance between the stimuli and responses and hence takes into account the relatedness of the two. In addition to these two measures, average response time was also considered as another possible measure of comparing the test stimuli processed differently.

The experiments involving consonant-to-vowel intensity ratio (CVR) modification were divided into four test sets: CV9, VC9, CV6, and VC6. The CV9 and VC9 tests involved the three unvoiced stop consonants /p, t, k/ in the CV and VC contexts respectively of the three vowels /a, i, u/. The CV6 and VC6 tests involved the stop consonants /p, t, k, b, d, g/ in the CV and VC contexts respectively of the vowel /a/. Five versions of each stimulus were synthesized, one with no modification and the other four with CVR modification of +3, +6, +9, and +12 dB. The aim of the experiments was to study the effects of increasing CVR on consonant recognition in terms of the following: vowel-context, vowel impairments, consonant position in the syllable, and response times. In addition, the effect of CVR modification on the transmission of information with respect to vowel feature and features of consonantal place and voicing were also to be studied.

From the results of the CV9 and VC9 tests, it has been observed that increasing CVR does improve recognition scores in the presence of masking broadband noise for normal-hearing subjects. However, the recognition scores for stops in the syllable-final (VC) context are seen to be higher than those for stops in the syllable-initial (CV) position for all the CVRs. Similar results have been obtained for the CV6 and VC6 tests too.

The pattern of confusions in the identification of place of articulation remains generally unaffected for all the CVRMs in all the four tests. Since increase in CVR has been seen to be more effective in bringing down the overall level of confusions in the /VC/ context as compared to that in the /CV/ context, it is suggested that increasing CVRM suppresses forward masking of the consonant by the vowel. more effectively than backward masking.

Increasing CVR can sometimes result in vowel confusions as seen from the /ip/-/up confusions which were found to increase in severity for higher CVRs. This confusion was found to increase from 23% at 6 dB CVR modification to 27% at 9 dB CVR modification and 34% at 12 dB CVR modification. Thus, it appears that the variability of vowel confusions is another factor that could set a limit on the amount of CVR modification that may be used. In experiments with closed-set stimuli, this limit as well as the specific vowel confusions may depend upon the test set involved

Information transmission analysis results for the CV9 and VC9 tests show that both overall information transmitted as well as transmission of consonant feature increases appreciably with increasing CVR. However, the overall information transmitted as well as transmission of consonant feature is seen to be superior in the VC context as compared to those in the CV context. The information transmission of vowel feature for the two tests is seen to be near-perfect for both the no-noise and 12 dB SNR cases and slightly lower for the 6 dB SNR case

Information transmission analysis results for the CV6 and VC6 tests show near-perfect transmission of overall information as well as information transmitted with respect to place and voicing features for the no-noise case. For the 12 dB and 6 dB SNR cases, the information transmitted with respect to all the three features is found to increase appreciably with increase in CVR.

The results of the information transmission analysis for CV6 and VC6 tests further reveal that the information transmitted about place feature is superior in the VC context over that in the CV context. However, the transmission of voicing information is superior in the CV context over that in the VC context.

The average response times for all the four tests do not appear to show any deterioration and in most cases appear to indicate a quicker response by the subjects for increasing values of CVR

These experiments were done to study the effect of increasing CVR on stop consonants from three different view-points: recognition score, amount of information transmitted on the basis of feature-classification, and the average response time. It is concluded that increasing CVR has a positive impact in all the cases. A CVR modification of upto about 10 dB may be used without any adverse effect on vowel recognition for the test stimuli considered here.

The experiments involving consonant duration modification were performed to study the effect on perception of altering the duration of each of the acoustic segments that constitute the consonant phoneme. Stop consonants /p, t, k, b, d, g/ in the CV context of the vowel /ɑ/ were chosen as the stimuli to separately consider the effect of modifying the following: burst duration (BD), formant transition duration (FTD), and voice onset time (VOT).

From the results of the BD modification test, it has been observed that in the case of no masking noise, the recognition score was unaffected when burst duration was doubled. For the 12 dB and 6 dB SNR presentations, a very slight increase (1%) in recognition score was observed when burst duration was doubled. From the results for the no masking noise and 12 dB cases in the FTD modification test, the recognition score was found to decrease slightly as the transition duration was doubled. However, for the 6 dB SNR case, the score was found to increase by about 4% for 50% FTD increase and then decrease by 4% for 100% FTD increase. For the VOT modification test, a generally decreasing trend in scores was observed for all the three noise conditions as VOT was increased.

Information transmission analysis results for the BD modification test show that while the overall information transmitted decreases slightly (2%) when burst duration is doubled in the no masking noise case, it remains unaffected in the 12 dB case, and increases slightly (2%) in the 6 dB case. The transmission of place feature is found to increase slightly when burst duration is doubled for all the noise cases. However, the transmission of voicing feature information is found to decrease with increasing BD.

Information transmission analysis results for the FTD modification test stimuli show a 4% decrease in overall information transmitted when transition duration is doubled in the no masking noise case. However, for the 12 dB SNR case, an increase of 3% is observed. For the 6 dB SNR case, the overall information transmitted is seen to increase appreciably (8%) for 50% FTD increase and then decrease slightly (2%) for 100% FTD increase. The transmission of place information is also seen to follow a similar trend. The transmission of voicing information is perfect for all the FTD modifications in the no masking noise case. However, for the 12 dB and 6 dB SNR cases, the transmission of voicing information is seen to decrease considerably by 27% and 25% respectively, when transition duration is doubled.

Information transmission analysis results for the VOT modification test show a generally decreasing trend in overall information transmitted as well as transmission of place and voicing information for all the noise conditions when VOT is increased.

The average response time was found to generally decrease (improve) with increasing duration modification for the BD and FTD modification tests. However, for the VOT modification test, the response time was found to increase with increasing VOT.

The above results suggest that of the three acoustic segments considered here of consonantal duration, increase in FTD yields the most positive results. An FTD increase of upto 50% is seen to improve the performance. At lower SNRs, this amount of FTD modification may be combined with burst duration modification and expected to yield better performance. However, VOT does not appear to be a suitable parameter for modification because of the reduction in performance seen with increasing VOT.

The present study has established that C/V intensity ratio modification and, to a lesser extent, modification in formant transition duration would improve speech perception. This study was done using nonsense CV and VC syllables as test material and subjects with simulated sensorineural impairment. In order to further evaluate the effectiveness of this approach, work should be carried out on building a speech processor and conducting experiments on subjects with hearing impairment.

In the earlier studies reported, the modifications have been done on natural speech by visual inspection of the waveform, intensity envelope, and spectrogram and introducing the modifications. For instance, the consonant duration was increased by merely replicating small segments within the consonant. This cannot be expected to reflect the actual changes associated with acoustic segments. What is called for is some method whereby the effects, on perception, of increasing the duration of each acoustic segment independently can be monitored.

In this study, the parameters were modified at the synthesis stage. In building a speech processor, an analysis/synthesis approach would possibly have to be adopted in which the subsegment boundaries can be identified and the subsegments resynthesized after appropriate modifications. It is obvious that this analysis/synthesis approach would require a processing time delay extending over several subsegment durations.

It has been previously shown that for speech processing for cochlear prostheses, delays of upto about 120 ms in the speech processing and stimulus encoding should not interfere with the benefits of auditory signal in audiovisual comprehension of connected speech. Hence, in the present case, the segmentation, feature classification, and speech enhancement processes should be completed within this time.

TG Thomas, Experimental Evaluation of Improvement in Speech Perception with Consonantal Intensity and Duration Modification, Ph. D. Thesis, Department of Electrical Engineering, Indian Institute of Technology, Bombay, 1996.

ABSTRACT

A promising scheme for enhancing the speech signal is based on studies of speaking clearly for the hearing impaired. Studies of the differences between "clear" speech and "conversational" speech have identified certain consistent acoustic modifications of the speech signal in clear speech. Two characteristics of clear (intelligible) speech, namely consonant-to-vowel intensity ratio (CVR) and consonant duration (CD), have been chosen for evaluation in the present investigation.

The aim of the experiments reported here was to study the extent of CVR modification and CD modification that would be helpful in speech perception by the hearing impaired. A computerized test administration system was developed in order to automate the experiments. It was decided to use synthetic stimuli since the segmentation of individual phonemes would be accurate, and it would be possible to manipulate the various acoustic segments independently. The stimuli involved stop consonants in the consonant-vowel (CV) and vowel-consonant (VC) contexts of various vowels. The subjects were normal-hearing persons and hearing impairment was simulated by mixing each stimulus with masking broadband noise. The results of each experimental run were obtained in the form of a stimulus-response confusion matrix. The confusion matrices were evaluated on the basis of recognition scores and information transmission analysis. Average response time was also considered as another possible measure of comparing the test stimuli processed differently.

The results show that increasing CVR does improve recognition scores. Some vowel confusions were observed in the VC context at higher CVRs, but as long as CVR modification was restricted to about 9-10 dB there was no adverse effect on the recognition of vowels. The information transmitted on the basis of feature-classification, as well as the average response times were also found to improve with increasing CVR.

For the CD modification experiments, the acoustic segments that constitute the consonant phonemes were altered in duration separately to study their individual effects. The results suggest that at higher noise levels, a formant transition duration modification of upto about 50% may be combined with burst duration modification and expected to yield better performance. Voice onset time (VOT) does not appear to be a suitable parameter for modification as performance decreased with increasing VOT.

These results suggest that it would be fruitful to develop a scheme that can identify the boundaries of the various phonemic segments and subsegments and then perform appropriate modifications in C/V ratio, formant transition duration, and burst duration. This scheme could then be tested on the hearing impaired.

KEY WORDS: *Speech enhancement, Speech perception, Hearing impaired, Consonant-to-vowel intensity ratio, Consonant duration*

LIST OF COURSES/SEMINAR COMPLETED

S. No.	Course No.	Course Name	Course Credit
1.	EE3603	Digital Signal Processing and its Applications	6.0
2.	EE3629	Biomedical Instrumentation	6.0
3.	EE3612	Telematics	6.0
4.	EE3698	Special Topics in Electrical Engineering	6.0
5.	EES801	Seminar	6.0
		Total	30.0