ENHANCEMENT OF SPEECH INTELLIGIBILITY USING ACOUSTIC PROPERTIES OF CLEAR SPEECH

Thesis
submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

by

A. R. Jayan
(Roll No. 05407303)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay

2014

Dedicated to my mother

Indian Institute of Technology Bombay
Department of Electrical Engineering

Ph.D. Thesis Approval

Thesis entitled "**Enhancement of Speech Intelligibility Using Acoustic Properties of Clear Speech**" by **A. R. Jayan** is approved for the award of the degree of **Doctor of Philosophy.**

Supervisor:	P.C. Pandey	(Prof. P. C. Pandey)
Internal Examiner:	Preeti Rao	(Prof. Preeti Rao)
External Examiner:	A. Kumar 26/5/2014	(Prof. Arun Kumar)
Chairman:	26/5/14	(Prof. N. Prabhu)

Date: 26th May, 2014
Place: Mumbai

Indian Institute of Technology Bombay

CERTIFICATE OF COURSE WORK

This is to certify that Mr. A. R. Jayan (Roll No. 05407303) was admitted to the candidacy of Ph.D. Degree in July 2006, after successfully completing all the courses required for the Ph.D. Degree Programme. The details of the course work done are given below.

Sr. No.	Course No.	Course Name	Credits
1	EE603	Digital Signal Processing and its Applications	6
2	EE616	Electronic Systems Design	6
3	EE679	Speech Processing	6
4	EES801	Seminar	4
5	EE712	Embedded Systems Design	6
6	HS699	Communication and Presentation Skills	
		Total Credits	**28**

I.I.T. Bombay
Dated:

Dy. Register (Academic)

A. R. Jayan / Supervisor: Prof. P. C. Pandey, “Enhancement of speech intelligibility using acoustic properties of clear speech,” *Ph.D. Thesis*, Department of Electrical Engineering, Indian Institute of Technology Bombay, May 2014.

Abstract

Speech intelligibility can be enhanced using acoustic properties of “clear speech”, the speech produced by a speaker with an intention to improve intelligibility in a difficult communication environment. The research objective is to devise a signal processing technique based on the properties of clear speech for improving perception of stop consonants for use in speech communication devices and hearing aids. The method assumes clean speech to be available and processing is performed to make it robust towards further degradations under adverse listening conditions.

Modification of speech characteristics around acoustic landmarks, the regions with a concentration of acoustic cues, is expected to improve speech intelligibility. Detection of landmarks associated with stop consonants was investigated using (i) subband energies and centroids, (ii) parameters from Gaussian mixture modeling, (iii) spectral moments, and (iv) spectral moments with tone-addition. Comparing the algorithmic and computational delays involved in landmark detection, rate of change of spectral centroid derived from tone-added speech spectrum was identified as the parameter most suited for real-time detection of burst onset landmarks. Automated enhancement of speech intelligibility by consonant-vowel ratio (CVR) modification and time-scale modification was investigated. CVR modification involved amplification of vowel-to-consonant (VC) and consonant-to-vowel (CV) transition segments by 9 dB. Sinusoidal model based analysis-synthesis was used for time-scale modification of CV transition segment by a factor of 1.5. Listening tests were conducted on normal-hearing subjects using isolated VCV utterances with speech-spectrum shaped noise as a masker. CVR modification improved recognition scores by 7, 18, and 25% at SNRs of 0, −6, and −12 dB, respectively. No statistically significant improvements were obtained for time-scale modification.

Further investigations were performed on CVR modification using a real-time compatible algorithm. Test material involved utterances with CV syllables, VC syllables, and MRT wordlists. Listening tests were conducted using normal-hearing subjects with speech-spectrum shaped noise as the masker. The improvements in recognition scores for the CV syllables were 8, 9, and 19% at SNRs of 0, −6, and −12 dB, respectively. The corresponding improvements were 9, 11, and 14% for VC syllables. For MRT wordlists, the corresponding improvements were and 8, 9, and 11%, and the improvements were equivalent to an SNR advantage of 3 dB. There was no significant increase in the response times for the processed stimuli indicating no significant increase in the perceptual load. The technique for CVR modification was implemented on a DSP board based on a 16-bit fixed point processor with on-chip FFT hardware and tested for satisfactory real-time operation. Thus the investigations have shown that CVR modification using the proposed technique may be used for improving speech perception under adverse listening conditions.

CONTENTS

List of Figures

List of Tables

List of Symbols and Abbreviations

Symbols

$A_g(n)$	amplitude of Gaussian component *g* for frame *n*
A_l^k	amplitude of l^{th} sinusoid in frame *k*
A_m	maximum gain
$B_g(n)$	bandwidth of Gaussian component *g* for frame *n*
$dF_c(n)$	rate of change of spectral centroid for frame *n*
$dF_{ct}(n)$	rate of change of spectral centroid from tone added spectrum for frame *n*
$E(n)$	energy of frame *n*
$E_b(n)$	subband energy parameter in the band *b* in dB
$E_b'(n)$	first difference of subband energy
$E_p(n)$	peak energy of frame *n*
$E_s(n)$	smoothed energy of frame *n*
$f_b(n)$	subband centroid in Hz
$f_c'(n)$	first difference of subband centriod
f_s	sampling frequency
$F_c(n)$	spectral centroid for frame *n*
$F_{ct}(n)$	spectral centroid for tone-added frame *n*
$F_k(n)$	spectral kurtosis
$F_{kt}(n)$	spectral kurtosis of tone-added spectrum
$F_m(n)$	m^{th} spectral moment for frame *n*
$F_s(n)$	spectral skewness
$F_{st}(n)$	spectral skewness of tone-added spectrum
$F_\sigma(n)$	spectral standard deviation
$F_{\sigma t}(n)$	spectral standard deviation of tone-added spectrum
$G(n)$	gain for frame *n*
$G_m(n)$	maximum gain for frame *n*
k	DFT bin index
K	time step
L_k	number of sinusoidal tracks in frame *k*
n	frame index
N	frame length in samples
$p(g\|k)$	probability of frequency measurement *k* to come from Gaussian *g*
$p(n,k)$	normalized spectrum
$r_A(n)$	rate of change based on Gaussian amplitudes
$r_c(n)$	rate of change based on Gaussian parameters
$r_\mu(n)$	rate of change based on Gaussian means
$r_\sigma(n)$	rate of change based on Gaussian variances
$s(n)$	analysis frame
$s'(n)$	synthesized frame
$S_n(k)$	median smoothed log magnitude spectrum
$\hat{S}_n(k)$	Gaussian approximation of smoothed log magnitude spectrum
t_a	analysis time instants
t_s	syntheis time instants
$T_{rE}(n)$	transition index based on band energy parameters
$T_{rEC}(n)$	transition index based on band energies and centroids
$X(n,k)$	DFT of frame *n*
$\|X(n,k)\|$	magnitude spectrum of frame *n*
$X_t(n,k)$	DFT of tone-added frame *n*

$\mathbf{y}(n)$	parameter set for frame *n*
$w_{gn}(n)$	mixture weight of Gaussian component *g*
α	decay time instants
β	scaling factor for time-scale modification
γ	gain scaling factor
θ_h	upper threshold frequency
θ_l	lower threshold frequency
θ_l^k	phase of l^{th} sinusoid in frame *k*
$\mu_{gn}(n)$	mean of Gaussian component *g*
$\sigma_{gn}(n)$	standard deviation of Gaussian component *g*
ϕ_l^k	phase offset of l^{th} sinusoid in frame *k*
φ_l^k	unwrapped phase of l^{th} sinusoid in frame *k*
ω_l^k	angular frequency of l^{th} sinusoid in frame *k*

Abbreviations

C	intensity modification of consonant segment
CE	consonant enhancement
CF	intensity and spectral modification of consonant segments
CFT	intensity and spectral modification of consonant segments along with intensity modification of transition segments
CT	intensity modification of consonant and transition segments
CV	consonant-vowel syllable
CV6	listening test involving 6 CV utterances
CV9	listening test involving 9 CV utterances
CVC	consonant-vowel-consonant syllable
CVR	consonant-to-vowel ratio
dB	decibel
DFT	discrete Fourier transform
DMA	dynamic memory access
EC	energy and centroid
EM	expectation maximization
ERB	equivalent rectangular bandwidth
F_0	fundamental frequency
F1	first formant
F2	second formant
F3	third formant
F4	fourth formant
FFT	fast Fourier transform
GMM	Gaussian mixture model
HMM	hidden Markov model
kHz	kilo Herz
MB	mega byte
MRT	Modified Rhyme Test
PSOLA	pitch synchronous overlap add
RMS	root mean square
ROC	rate of change
ROC-MD	Mahalanobis distance based rate of change
ROR	rate of rise
SFM	spectral flatness measure
SM	spectral moments
SMTA	spectral moments with tone-addition

SNR	signal-to-noise ratio
SPL	sound pressure level
SUS	semantically unpredictable sentences
SVF	spectral variation function
TD-PSOLA	time-domain pitch-synchronous overlap add
TI	Texas Instruments
TIMIT	Texas Instruments/MIT speech corpus
VC	vowel-consonant syllable
VC6	listening test involving 6 VC utterances
VC9	listening test involving 9 VC utterances
VCV	vowel-consonant-vowel syllable
VOT	voice onset time

DECLARATION

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

A. R. Jayan
(Roll No: 05407303)

Date: 27/05/2014

Chapter 1

INTRODUCTION

1.1 Problem overview

Speech intelligibility can be enhanced by modifying the speech characteristics in selected regions so as to make speech more robust towards subsequent degradations. The motivation for this approach is derived from "clear speech", the speech produced by a speaker with an intention to improve intelligibility in a difficult communication environment (Chen, 1980; Picheny et al., 1985). Reduced speaking rate with frequent and lengthy pauses, increased intensity and duration of consonant segments, clearly released and more intense stop release bursts, higher fundamental frequency with increased dynamic range, and more targeted vowel formants are some of the important acoustic characteristics of clear speech (Chen, 1980; Picheny et al., 1985; Picheny et al., 1986). Clear speech is more intelligible than conversational speech for normal-hearing listeners in adverse listening conditions, hearing-impaired listeners, children with learning disabilities, and non-native listeners (Payton et al., 1994; Bradlow and Bent, 2002; Bradlow et al., 2003).

Over the last three decades, significant research has been carried out to identify the acoustic properties of clear speech responsible for its increased intelligibility and to incorporate some of these properties in conversational speech to improve its intelligibility (Gordon-Salant, 1986, 1987; Montgomery and Edge, 1988; Picheny et al., 1989; Thomas, 1996; Kennedy et al., 1997; Vaughan et al., 2002). Some researchers have used a perceptual approach in which the modification of the speech characteristics is restricted around the regions with a concentration of acoustic cues (Hazan and Simpson, 1998; Colotte and Laprie, 2000; Ortega et al., 2000; Skowronski and Harris, 2006; Yoo et al., 2007; Tantibundhit et al., 2009). These regions are known as "landmarks" (Stevens, 1981; Stevens et al., 1992; Liu, 1996; Park, 2008) and are generally associated with spectral transitions. Real-time detection of the landmarks of interest and modification of the selected speech characteristics around these regions without introducing perceptible artifacts are the main challenges in the practical

application of speech intelligibility enhancement techniques based on the acoustic properties of clear speech.

1.2 Research objective

The objective of the present investigation is to devise a technique suitable for real-time implementation in speech communication devices and hearing aids to improve speech intelligibility. Investigations are performed on signal processing techniques for modification of specific speech segments with algorithmic and computational delays compatible with real-time processing. The method assumes clean speech to be available and processing is performed to make it robust towards further degradations under adverse listening conditions. The effectiveness of the method is evaluated under different levels of masking noise using normal-hearing subjects.

1.3 Thesis outline

The second chapter presents a brief review of the speech intelligibility enhancement techniques using acoustic properties of clear speech and the description of the proposed scheme for intelligibility enhancement. Investigations on landmark detection techniques for application in speech intelligibility enhancement are presented along with performance evaluations in the third chapter. Signal processing for performing consonant-vowel ratio (CVR) modification and time-scale modification around the detected landmarks for enhancement of speech intelligibility along with the results of experimental evaluation using normal hearing subjects are presented in the fourth chapter. In the next chapter, a method for real-time CVR modification of continuous speech along with the implementation details and results of experimental evaluations are presented. Summary of the investigations, conclusions derived, and some suggestions for future work are presented in the last chapter.

Chapter 2

ENHANCEMENT OF SPEECH INTELLIGIBILITY USING ACOUSTIC PROPERTIES OF CLEAR SPEECH: A REVIEW

2.1 Introduction

This chapter provides a brief description of the salient acoustic properties of clear speech, a review of speech intelligibility enhancement techniques using the properties of clear speech, and the motivation for the present investigation.

2.2 Acoustic properties of clear speech

A talker in a difficult communication environment (such as a noisy background or while talking to a hearing-impaired listener) usually alters the speaking style to make speech more intelligible. The speech produced in this mode of speaking is known as “clear speech”. In comparison with conversational speech, clear speech is more intelligible for normal-hearing listeners under adverse listening conditions, hearing-impaired listeners, children with learning disabilities, and non-native listeners (Picheny et al., 1985; Payton et al., 1994; Bradlow and Bent, 2002; Bradlow et al., 2003). Figure 2.1 shows waveforms and spectrograms of a sentence spoken in conversational and clear modes by a speaker. There is a reduction in speaking rate and an increase in the intensity of stop release bursts. The acoustic properties of clear speech that mainly contribute towards its increased intelligibility have been investigated by many researchers.

Acoustic properties of conversational and clear speech were compared at the segmental level by Chen (1980) using CV syllables involving stop consonants /b, d, g, p, t, k/ and vowels /a, i, u/. In clear speech, there was non-uniform increase in the stop closure duration, voice onset time, formant transition duration, and vowel duration. The formant frequencies were more targeted in clear speech, which resulted in a spread-out vowel triangle with tightly packed formant distribution at the vertices. There was a noticeable increase in fundamental frequency indicating a raising of the larynx during clear speech production.

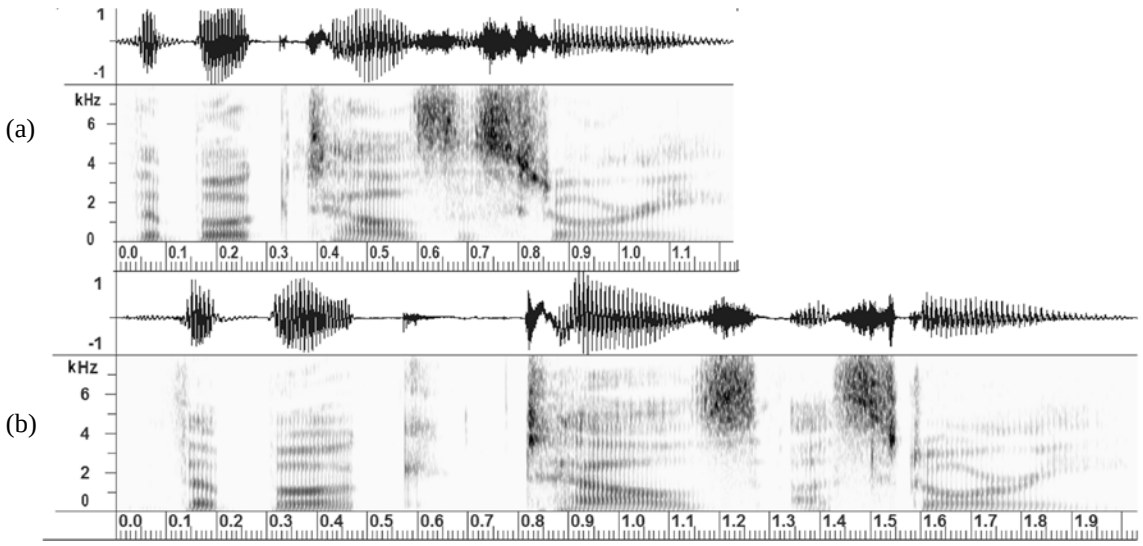


Figure 2.1 Waveforms and spectrograms of the utterance "The book tells a story" in (a) conversational and (b) clear mode (audio source: http://www.acoustics.org/press/145th/clr-spch-tab.htm).

Picheny et al. (1986) compared acoustic properties of clear speech and conversational speech at the sentence, word, and phoneme levels. A set of fifty nonsense sentences spoken in clear and conversational mode from three speakers was used in the study. At the sentence level, the speaking rate nearly halved for clear speech due to more frequent and lengthy pauses. The deletion of release bursts particularly for stops in the word-final position was less frequent in clear speech. Compared with conversational speech, a larger number of sound insertions occurred after word-final nasals, voiced stops, and fricatives. Significant aspiration was observed for word-final unvoiced sounds indicating continuation of vocal effort even after the consonant release. The phoneme durations, segmental power, vowel formant frequencies, and short-time spectra were compared at the phonemic level. Context and phoneme dependent non-uniform increase in phoneme duration was observed in clear speech. There was a significant increase in the overall duration of stop consonants due to increased durations of closure, frication, aspiration, and voice onset time. The ratio of RMS value of the consonant segment to that of the nearby vowel segment, referred to as the consonant-vowel ratio (CVR) was considerably higher in the clear speech. Formant frequencies approached the target values more frequently and there was an upward shift in the intensities and locations of spectral peaks in the short-time spectrum.

2.3 Intelligibility enhancement using properties of clear speech

The intelligibility enhancement techniques using the acoustic properties of clear speech can

be broadly classified into duration modification and intensity modification. A brief review of some of these techniques is given in the following subsections.

2.3.1 Techniques based on duration modification

Reduction in speaking rate is the most noticeable difference between conversational and clear speech. Picheny et al. (1989) evaluated the effect of reduction of speaking rate on speech intelligibility. The speaking rates of conversational and clear speech were equated by uniform time-scale modification. Malah's algorithm (Malah, 1979) which interpolated or decimated short-time speech spectrum was used for performing time-scale modification. Testing was performed on five hearing-impaired subjects using nonsense sentences as the test material. Test results showed time-scale modified speech to be less intelligible compared to the unprocessed speech. Gordon-Salant (1986) evaluated the effect of consonant duration modification on speech perception using manually annotated CV syllables. The CV boundary was located manually by audio and visual inspections. Consonant duration was doubled by duplicating waveform corresponding to each pitch period for voiced consonants and every 20 ms segment for unvoiced consonants. Results of listening tests showed only marginal improvements in recognition scores for stimuli with modified consonant duration.

Thomas (1996) investigated the effect of consonant duration modification on speech perception using synthetic CV syllables. Six stop consonants (/b, d, g, p, t, k/) in the context of vowel /a/ were used as the test material. The effects of increasing burst duration (by 100%), formant transition duration (by 50% and 100%) and voice onset time (by 50% and 100%) were separately investigated. Overall duration of the syllable was maintained constant by adjusting the duration of the vowel segment. Evaluations were performed on four normal-hearing subjects with stimuli mixed with broadband noise at different SNRs. Expansion of formant transition duration and voice onset time reduced the consonant recognition scores whereas marginal improvements were observed for stimuli with modified burst duration.

Uchanski et al. (1996) investigated the effect of non-uniform time-scale modification at the segmental level on speech intelligibility. Griffin and Lim's algorithm (Griffin and Lim, 1984) which estimated the signal from time-scaled version of the magnitude spectrum using an iterative technique was used for performing time-scale modification. The regions for modification were located manually. Intelligibility of time-scale expanded conversational speech (with segment durations equated to clear speech) and time-compressed clear speech (with segment durations equated to conversational speech) were compared with that of naturally produced conversational and clear speech. Listening tests involved the identification of keywords in nonsense carrier sentences. Tests were performed on five normal-hearing subjects with speech mixed with white noise at an SNR of −4 dB and on four subjects with sensorineural hearing loss. The naturally produced clear speech was more intelligible than the

unprocessed conversational speech by nearly 15%. For normal-hearing and hearing-impaired subjects, time-scale modified speech was less intelligible compared to the unprocessed speech. For normal-hearing subjects, time-scale expansion of conversational speech decreased the scores by 5% and time-compression of clear speech decreased the scores by 24%, when compared with the corresponding unprocessed speech. The recognition score for the time-compressed clear speech was 9% less than that of the unprocessed conversational speech. The same trend was observed for hearing-impaired subjects.

The effect of insertion and deletion of pauses on speech intelligibility has also been investigated (Picheny et al., 1989; Uchanski et al., 1996) using unprocessed conversational and clear speech, pause inserted conversational speech, and pause deleted clear speech. Deletion of pauses in clear speech reduced the word recognition scores by 4%. Insertion of pauses in conversational speech also reduced the scores by 8%. These studies indicated the reduction in speaking rate not to be a major contributor towards the intelligibility advantage of clear speech.

Vaughan et al. (2002) investigated the effect of duration modification of unvoiced consonants on speech recognition by two groups of normal-hearing subjects (young with average age of 26 years, older with average age of 67 years) and a group of hearing-impaired subjects (average age of 70 years) with mild-to-moderate sensorineural loss. The slowed down speech was expected to compensate for the age related changes in the cognition process for continuous speech. Based on variation of energy parameters, speech was segmented into silence, voiced and unvoiced consonant, and vowel regions. Time scaling was performed by duplicating short-time segments of the speech signal. Time-expansion factors of 1.2 and 1.4 were used during unvoiced stop consonants and fricatives, and other segments were left unmodified. Listening tests were conducted using paragraphs consisting of 10 sentences each, presented in 12-talker babble noise background at an SNR of 4 dB. Recognition scores calculated for keywords in the paragraphs were lower for the time-scale modified speech than for the unprocessed speech for all groups of subjects in quiet as well as in the presence of noise. Scores for time-scaling factor of 1.4 were slightly lower than those for 1.2 in quiet, but higher in the presence of noise.

Intelligibility of clear speech produced at normal conversational speaking rates by trained speakers was investigated by Krause and Braida (2002). Listening tests were conducted on eight normal-hearing subjects using nonsense sentences mixed with speech-shaped noise at an SNR of -1.8 dB. Clear speech produced at the conversational speaking rate was nearly 14% more intelligible than the conversational speech. An examination of the acoustic properties of conversational and clear speech produced at normal speaking rates (Krause and Braida, 2004) showed that the long-term spectral power above 1 kHz was more in clear speech than in conversational speech. Increased modulation depth of the temporal

envelope, increased average value and dynamic range of fundamental frequency, and increased energy near second and third formants in the short-time spectrum were identified as the major contributors to the intelligibility advantage of clear speech.

Liu and Zeng (2006) investigated the effect of altering the speaking rate by uniform time-scale expansion, time-compression, and insertion of pauses. The contribution of temporal properties (envelope, periodicity, and fine structure) on speech intelligibility was also investigated. Three experiments were conducted using 144 sentences in clear and conversational speaking styles. In the first experiment, conversational speech was uniformly expanded to have duration of clear speech, and clear speech was uniformly time-compressed to have duration of conversational speech. Time-scaling was performed using pitch-synchronous overlap and add (PSOLA) method in which original signal was decomposed into pitch cycles centered at the pitch synchronous marks. These cycles were duplicated or deleted based on the time-scaling factor to get time-expanded or time-compressed speech, respectively. In the second experiment, speaking rate of the conversational speech was equated with that of the clear speech by proportionately increasing the silent intervals between phonetic segments in conversational speech. Silent gaps shorter than 10 ms were kept intact to avoid alteration of stop closures during voiced stop consonants. This method introduced minimal signal processing artifacts. In the third experiment, fine structures of conversational and clear speech were transformed to each other by signal processing. The stimuli were band-pass filtered using 16 logarithmically spaced filters and decomposed to envelope and fine structure by Hilbert transform. Non-uniformly stretched envelope of the conversational speech was used to amplitude modulate the fine structure of the clear speech. Similarly, the non-uniformly compressed envelope of the clear speech was used to amplitude modulate the fine structure of the conversational speech. Testing was performed on 10 normal-hearing subjects using speech mixed with speech-spectrum shaped noise, at SNR ranging from −15 to 10 dB. The results of the first experiment showed that the slowed down conversational speech was less intelligible than the natural conversational speech by nearly 10%, and more intelligible than time-compressed clear speech by nearly 20% at −5 dB SNR. The second experiment showed that the silence insertion resulted in nearly 9% improvement in recognition scores. The third experiment revealed the complementary contributions of fine structure and temporal envelope to intelligibility. At SNRs below −5 dB, fine structure contributed more, whereas at quiet and positive SNRs, envelope was found to be more important for intelligibility.

The acoustic differences between characteristics of clear and conversational speech are distributed in a segment-specific and non-uniform fashion. This makes automated transformation of conversational speech to clear speech by signal processing nearly impossible. Some of the temporal properties identified to be contributing towards the

intelligibility advantage of clear speech cannot be introduced in an automated fashion due to the signal processing difficulties. The feasible signal processing solutions have not resulted in a significant intelligibility advantage, possibly because the well defined formant targets achieved during clear speech production may not be achieved by mere time-scale modification of conversational speech.

2.3.2 Techniques based on consonant intensity modification

The ratio of RMS energy of consonant segment relative to the nearby vowel segment is called the consonant-vowel ratio (CVR) (Gordon-Salant, 1986; Freyman and Nerbonne, 1989). CVR depends on many factors including consonant type, vowel context, and speaking style. Gordon-Salant (1986) reported an average CVR of −9.86 dB for CV syllables involving 19 consonants (/b, d, g, p, t, k, m, n, s, z, v, f, w, j, l, r, ʃ, θ, ð/) paired with 3 vowels (/a, i, u/). CVRs for nasals, glides, and liquids were higher than the average CVR. For fricatives, CVRs were lower than the average CVR. Kennedy et al. (1997) reported CVRs for nonsense VC syllables to be in the range of −15 to −28 dB for unvoiced stops and fricatives (/p, t, k, f, ʃ, s, θ/), −10 to −18 dB for voiced stops and fricatives (/b, d, g, v, ð, z/), and −4 to −9 dB for nasals (/m, n, ŋ/).

Guelke (1987) reported a technique for improving identification of stop consonants (/b, d, g, p, t, k/) by automatically detecting and amplifying stop release bursts. Intensity enhancements by 9, 15, and 17 dB were applied for burst enhancement durations of 14, 17, 24, and 40 ms. Stimuli were mixed with masking noise and the level of the noise was adjusted to get 50% recognition score for the stimuli without enhancement. Listening tests in the presence of masking noise showed that the recognition scores increased from 51% with no enhancement to 90% for stimuli with stop release bursts enhanced by 17 dB for an enhancement duration of 24 ms. Gordon-Salant (1986; 1987) compared the performance of normal-hearing subjects and subjects with sensorineural loss on identification of intensity and duration altered CV syllables. Four types of stimuli: unprocessed, CVR modified, duration modified, and combined CVR-duration modified, consisting of 19 consonants (/b, d, g, p, t, k, m, n, s, z, v, f, w, j, l, r, ʃ, θ, ð/) paired with three vowels (/a, i, u/) were used in 12-talker babble noise background. The regions for modification were identified manually. Listening tests were conducted at presentation levels of 75 and 90 dB SPL and 6 dB SNR. CVR was modifed by 10 dB and consonant duration was uniformly expanded by 100% by duplication of segments. CVR modification and CVR-duration modification improved the recognition scores. The improvement in recognition scores for normal-hearing subjects was 13% at the presentation level of 75 dB SPL and 9% at 90 dB SPL. The corresponding values were 16% and 11% for hearing-impaired subjects. The acoustic modification was found to have

significant interaction with the vowel context, and larger improvements were observed for vowel contexts of /i/ and /u/ than for /a/. Analysis of consonant confusions in terms of identifying consonants differing in place (front, middle, back), manner (glides, nasals, plosives, fricatives), and voicing (voiced, unvoiced) were performed. The CVR modification and CVR-duration modification improved the place, manner, and voicing recognition scores. The highest improvement in place recognition score was observed for CVR modified stimuli involving middle consonants paired with vowel /i/. The recognition scores for the unvoiced consonants were higher compared with the recognition scores for voiced consonants.

In a similar investigation, Montgomery and Edge (1988) used 100 CVC syllables to evaluate the modification of CVR and consonant duration. In CVR modification, CVR was made nearly equal to 0 dB by equating the consonant and vowel levels. In duration modification, the overall duration of continuant consonants and VOT of stop consonants were increased by nearly 30 ms, along with time-compression of the vowel segment by 30 ms. The effect of combined CVR-duration modification was also investigated. The evaluations were performed using 20 subjects with bilateral sensorineural loss. While the duration modification did not result in improvement of recognition scores, CVR and CVR-duration modification resulted in 10.5% and 12.2% increase in recognition scores, respectively.

Freyman and Nerbonne (1989) conducted experiments by varying CVR of CV syllables from 10 speakers, using eight unvoiced consonants (/p, t, k, t͡ʃ, f, θ, s, ʃ/) paired with vowel /a/. The means and standard deviations of CVRs estimated for each speaker across the consonants and for each consonant across the speakers varied widely from consonant to consonant. Out of the consonants, the standard deviation was lesser for /t, k/ (mean = –14.9 dB, s.d. = 2.2 dB) compared to /p/ (mean = –8.9 dB, s.d. = 6.7 dB). Three modifications were carried out. In the first modification, vowel levels in utterances from each of the speakers were equated without altering the natural CVRs of speakers. The resulting stimuli had equal vowel levels but different consonant levels depending on their natural CVRs. In the second modification, CVRs were equated to that of the speaker with highest CVR for each CV syllable, by keeping vowel levels same as in the first modification. In the third modification, consonant levels were equated for each CV syllable across speakers, maintaining their natural CVRs by varying the vowel levels. Listening tests were conducted on 50 normal-hearing subjects by presenting the stimuli monaurally in white noise background at 0 dB SNR. The mean scores were 51.6%, 63.3%, and 72.5% for the first, second, and third modifications, respectively. The results showed that the speech intelligibility was less correlated to the value of CVR, but more to the level of the consonant, and thus to their audibility.

Thomas (1996) evaluated improvement in speech perception by CVR modification using synthetic CV and VC syllables. Listening tests were conducted on 5 normal-hearing

subjects with 4 levels of CVR modification (+3, +6, +9, +12 dB). Broadband noise was used as the masker. Tests CV9 and VC9 involved 9 CV and VC syllables with unvoiced stop consonants (/p, t, k/) paired with 3 vowels (/a, i, u/). Tests CV6 and VC6 involved 6 CV and VC syllables with stop consonants (/b, d, g, p, t, k/) paired with vowel /a/. For CVR modification by +12 dB, CV9 test results indicated improvement in recognition scores from 62 to 88%, and from 52 to 80% at SNRs of 12 and 6 dB, respectively. The corresponding improvements in the VC9 tests were from 85 to 98%, and from 73 to 93%. For both CV and VC syllables, improvement was more in the /a/ and /u/ contexts than in the /i/ context. Results of CV6 tests indicated improvement in recognition scores from 81 to 94% and from 73 to 89% for CVR modification by 12 dB, at SNRs of 12 and 6 dB, respectively. The corresponding improvements in the VC6 tests were 85 to 97% and 71 to 94%, respectively. CVR modification was more effective in the VC context than in the CV context, indicating it to be more effective in reducing forward masking than backward masking.

Kennedy et al. (1997) investigated the CVRs required for maximizing consonant recognition and its dependency on consonant type, vowel environment, and listener's audiogram configuration. Consonant enhancement was performed by increasing the CVR by 0 to 24 dB in steps of 3 dB. The test material involved VC syllables consisting of 9 voiced consonants (/b, d, g, v, δ, z, m, n, η/) and 7 unvoiced consonants (/p, t, k, f, ʃ, s, θ/) paired with 3 vowels (/a, i, u/). Testing was conducted on 18 subjects with sensorineural loss, grouped in accordance with the audiogram configurations. Smooth curves were fitted on the recognition scores *vs* CVR modification, for each consonant in each vowel context, for each subject. Five different types of relations between recognition scores and CVR were identified. Recognition scores decreased with increase in CVR for nasals. The mean value of CVR modification for maximum improvement was 8.3 dB for voiced consonants (7.3, 9.5, 8.4 dB for /a/, /i/, /u/ contexts, respectively) and 10.7 dB for unvoiced consonants (9.8, 9.5, 12.5 dB for /a/, /i/, /u/ contexts, respectively). The improvements in consonant recognition scores ranged from 2 to 22% for unvoiced stops and fricatives and from 1 to 14% for voiced stops and fricatives.

Hazan and Simpson (1998) investigated cue-enhancement strategies for improving speech intelligibility. Thirty six manually annotated VCV syllables, comprising 12 consonants (/b, d, g, p, t, k, f, v, s, z, m, n/) and 3 vowels (/a, i, u/) were used as the test material. The regions involving vowel onsets and offsets, frication, nasal, stop release bursts, and aspiration were manually identified, and their RMS values were scaled during modification. Four processing approaches were used: modification of consonant intensity (C), modification of intensity of consonant and transition segments (CT), intensity and spectral modification of consonant segments (CF), and intensity and spectral modification of consonant segments along with intensity modification of transition segments (CFT).

Modification of consonant intensity (C) involved amplification of frication and nasal segments by 6 dB, and amplification of release burst in stop consonants by 12 dB. For modification of transition segments (T), five pitch cycles in the vowel segments adjacent to the consonant segment on either side were amplified by 6 dB. For spectral modification (F), burst segments were band-pass filtered to retain energy around 300 Hz for labials, 1.2 to 3 kHz for velars, and 2.5 to 4 kHz for alveolars. The bandwidths of the filters used were four times the auditory filter bandwidths at their center frequencies. To improve fricative identification, frication noise was concentrated above 1 kHz for /f, v/ and above 4 kHz for /s, z/ by filtering. Filtering was not performed for nasal consonants. Stimuli were combined with speech-spectrum shaped noise at SNRs of 0 and −5 dB. Listening tests on 13 normal-hearing subjects showed consistent improvements in recognition scores for all the processing conditions. Highest increase in recognition scores were obtained with the processing scheme CFT: 6% at 0 dB SNR and 12% at −5 dB SNR. Amplification of consonant segment (C) resulted in a significant improvement in intelligibility, while improvements gained from spectral modification (F) and modification of the transition segment (T) were marginal. Improvement was generally higher in the /a/ and /i/ contexts than in the /u/ context. Another experiment was conducted to evaluate the performance of the modification schemes in the presence of co-articulation. Semantically unpredictable sentences were used as the test material to reduce the effect of contextual information in the responses. Processing was carried out using manual annotation of the material. The bursts in sentence material were not filtered due to the difficulty in accurately estimating their center frequencies. Modified stimuli were combined with speech-shaped noise at 5 and 0 dB SNRs. The score at 0 dB SNR increased from 77.1 for unprocessed speech to 81.3% for CFT modified speech.

Li et al. (2010) investigated the effects of truncation, filtering, and noise masking on the recognition of stop consonants to find their perceptual cues in naturally produced speech. Listening tests using CV syllables with vowel /a/ were conducted on normal-hearing subjects. The labial stops were characterized by bursts in the 0.3 − 7.4 kHz band with 1 − 1.4 kHz peak, which got masked by white noise at 6 dB SNR. Perception of alveolar stops was mainly dependent on high frequency bursts above 3 − 4 kHz, which got masked by noise at 0 and −6 dB SNR, for /t/ and /d/, respectively. The velar stops were characterized by bursts in the mid-frequency band of 1.4 − 2 kHz, which got masked at 0 and −6 dB SNR for /k/ and /g/, respectively. The study showed that the stops often have conflicting cues and that robustness of a consonant to noise is related to the intensity of its dominant cue. In another study on recognition of fricatives in the presence of noise (Li et al., 2012), the perceptual cue was found to be contained in the frication noise. The lower edge of the band of frication and the amplitude modulation of the frication were found to be the most robust cues for place and

voicing, respectively. While the frication duration was not a robust cue for voicing, certain minimum durations were found to be necessary for perception of the fricatives. Truncation of fricative bursts beyond these durations, but not complete removal, resulted in confusions with plosives.

Kapoor and Allen (2012) investigated the effect of modifying the acoustic bursts on recognition of stop consonants. The effect of doubling, halving, and removing the release burst were investigated using CV syllables with stops /t, d, k, g/ and vowel /a/. Listening tests were conducted on normal-hearing subjects, with white noise used as a masker at SNRs ranging from −12 to +12 dB in 6 dB steps. There was an increase in recognition scores for the burst amplified stimuli and the shift in SNR for the same recognition score as that of the unmodified stimuli ranged from −5.6 to −4 dB. The recognition scores decreased for the burst attenuated stimuli and the shift in SNR for the same recognition scores as that of the unmodified stimuli ranged from 3.4 to 6.1 dB. The results indicated the intelligibility of stop consonants to be dependent on relative energy of the stop release bursts. The burst amplification did not change the pattern of confusions. The improvements in recognition scores were not very sensitive to slight errors in locating the burst boundaries.

In a recent study by Koning and Wouters (2012), the effect of enhancement of the onsets of the envelope of the speech signal on speech intelligibility in noisy conditions was studied using an eight-channel cochlear implant vocoder simulation. The processing involved splitting of the signal into eight frequency bands and emphasizing onsets of the speech envelope by deriving an additional peak signal at the onsets in each band. Selective amplification of the short duration onsets of the speech envelope did not cause a change in the perceived loudness of the speech material. Processing with the peak signal derived from the clean speech resulted in significant improvements in speech reception thresholds for stationary speech-shaped noise and a competitive talker as maskers. These results show that enhancement of the onsets in the speech envelope as part of signal processing in auditory prostheses has the potential of improving speech intelligibility.

2.4 Automated enhancement of speech intelligibility

Most of the intelligibility enhancement studies have used manual selection of regions and parameters for speech characteristics modification. For using such methods in a practical application, speech modification needs to be performed in real-time with automated selection of regions and parameters.

Nejime et al. (1996) reported a real-time system for slowing down fast speech to a comfortable rate, providing extra time for auditory memory processing. The speech signal was segmented into voiced, unvoiced, and silent frames using two intensity thresholds. Voiced segments were time-scale expanded using a time-domain pitch-synchronous method

for scaling factors of the order of 1.5. The unvoiced segments were kept unaltered and long duration silent segments were deleted. The incoming speech was buffered in memory during the processing, permitting continuous operation up to 3 minutes. Evaluations using 10 subjects with sensorineural loss showed no improvements in recognition scores for isolated word material. For sentence material, improvements of 1, 2, and 10% were observed for time-scaling factors of 1.25, 1.33, and 1.5, respectively. Listening tests using normal-hearing subjects with simulated cochlear loss (Nejime and Moore, 1998) showed a reduction in sentence recognition scores indicating that the processing for slowing down of the fast speech cannot help in reducing the adverse effects of cochlear loss although it may be helpful for persons with reduced speed of cognitive processing.

In an intelligibility enhancement technique reported by Colotte and Laprie (2000), a spectral variation function, based on mel-cepstral analysis, was used to locate the regions for enhancement. The spectral variation function detected 82% of manually located landmarks with a temporal accuracy of 20 ms. The processing involved time-scale modification and amplification of burst and fricative segments by +6 dB. Time-scale modification was performed using TD-PSOLA for scaling factors in the range of 1.8 to 2.0. Higher time-scaling factors resulted in audible distortions caused by transformation of the stops to fricatives. Listening tests were conducted using 13 normal-hearing subjects with 50 sentences from TIMIT database (Garofolo et al., 1993) as the test material. Half of the sentences were left unmodified and the others were processed by the algorithm. Listeners were asked to complete the missing words in the sentences. Recognition scores improved by 9% (from 72 to 81%) for intensity enhancement and by 14% (from 72 to 86%) for combined time and intensity enhancement. The improvements were statistically significant and were equally distributed among stops and fricatives.

Skowronski and Harris (2006) reported a real-time processing method for speech intelligibility enhancement based on redistribution of energy in voiced and unvoiced segments. Vowels, semivowels, nasals, voiced plosives, and voiced fricatives were attenuated and other low energy unvoiced sounds were amplified, maintaining the overall energy unaltered. A measure of spectral flatness derived from the short-time speech spectrum was used for voiced/unvoiced segmentation. A Schmitt trigger based thresholding was used to improve the performance of voiced/unvoiced detection. The errors in the classification of voiced and unvoiced frames were 2.9% and 18.8%, respectively. Listening tests were conducted using isolated utterances of confusable words from TI-46 corpus, in additive white noise at 0 and −10 dB SNRs. Speech from 16 speakers (8 male, 8 female) were tested on 25 subjects (9 native and 16 non-native speakers of English). The effects of processing and its dependencies on speaker effects and listener group (native *vs* non-native) were investigated. The enhancement improved intelligibility of speech from 9 speakers without degrading it for

speech from the other speakers. The effect of processing was found equally beneficial for native and non-native listeners. The sensitivity of the segmentation method to noise was identified by the authors as a limiting factor in the effectiveness of this enhancement scheme. The improvements in recognition scores were of the order of 5-10% at −10 dB SNR.

Yoo et al. (2007) decomposed speech signal to quasi steady-state and transient components. Voicing energy was removed by high-pass filtering and quasi steady-state component was identified using three time-varying band-pass filters based on a formant tracking algorithm. The transient part was obtained by subtracting the quasi steady-state part from the original speech. The transient part was amplified by different factors and added back to the quasi steady-state part to get the modified speech. Speech-spectrum shaped noise with steady spectrum from 100 Hz to 1 kHz and 12 dB/octave fall afterwards was used as the background noise. Modified Rhyme Test (MRT) was performed with stimuli presented at 6 different SNRs from −25 to 0 dB, with steps of 5 dB. In this experiment, SNR was defined as the ratio of formant energy to the noise energy, computed for each 10 ms frame. The test results indicated improvements of nearly 11, 18, 26, and 32% at SNRs of −10, −15, −20, and −25 dB, respectively.

Rasetshwane (2009) used a real-time algorithm to separate the speech into steady-state and transient components using a transitivity function based on rate of change of wavelet packet coefficients. Transient part of speech was extracted, modified by an amplification factor and then added to the steady-state part. Results of MRT with speech-spectrum shaped noise as the masker showed improvement in recognition scores of nearly 13% and 22% at −15 dB and −25 dB, with SNRs as defined by Yoo et al. (2007). The unprocessed stimuli performed better than the modified stimuli at SNRs above −5 dB. Tantibundhit et al. (2009) decomposed speech into steady-state and transient components using wavelet packet (Daubechies-16) decomposition. MRT was conducted on 14 normal-hearing subjects with speech added with speech-shaped noise at SNRs as defined by Yoo et al. (2007). The improvements in recognition scores were by 10, 0, 18, 27, 23 and 31%, for SNRs of 0, −6, −12, −18, −24, −30 dB, respectively.

Ortega et al. (2000) implemented a technique for identification of the regions to be enhanced in order to automate the cue-enhancement strategies proposed by Hazan and Simpson (1998). A broad-class HMM classifier was used for identifying the vocalic, fricative, nasal, stop, and silence regions. The detection rates for burst, frication, nasal, vowel onsets, and vowel offsets were 80, 75, 90, 78, and 76%, respectively. The corresponding insertion rates were 25, 9, 24, 16, and 17%, respectively. Testing was performed using 12 normal-hearing subjects with the speech signal mixed with speech-spectrum shaped noise at −5 dB SNR. The test material involved 80 monosyllabic words, 148 words from the lexical density

word test (Bradlow and Pisoni, 1999), and 24 nonsense VCV syllables. The processing improved recognition scores from 81 to 85% for monosyllabic words and 49 to 52% for lexical density words. There was a reduction in recognition scores from 74 to 72% for VCV syllables.

2.5 Summary

A review of earlier investigations indicates that CVR modification has the potential of improving consonant perception for normal-hearing listeners in noisy backgrounds and for hearing-impaired listeners. The effectiveness of these methods in speech intelligibility enhancement generally gets limited by the errors in landmark detection (involving deletions and insertions) and temporal inaccuracies (misalignment with actual landmarks). Further, the algorithmic and computational delays involved in many of these methods do not permit their implementation for real-time operation. In most of the investigations on intelligibility enhancement using acoustic properties of clear speech, clean speech is assumed to be available for processing and processing is performed to make it more robust under adverse listening conditions. In case of noisy speech input, processing for noise suppression using a method such as spectral subtraction (Loizou, 2007) has to precede the processing for CVR modification. We present investigations on techniques for landmark detection and processing for enhancing speech intelligibility under adverse listening conditions, with emphasis on low computational complexity and feasibility of real-time implementation in speech communication devices and hearing aids.

Chapter 3

LANDMARK DETECTION FOR SPEECH INTELLIGIBILITY ENHANCEMENT

3.1 Introduction

Landmarks are information rich regions in the speech waveform, where the acoustic cues essential for speech perception are supposed to be concentrated (Stevens, 1981; Stevens et al., 1992; Liu, 1996; Park, 2008). Landmarks are broadly classified into abrupt, non-abrupt, and vocalic. Acoustically abrupt landmarks coincide with the regions of major spectral changes and are produced by the movement of a primary articulator (lips, tongue blade, tongue body) or by a sudden glottal or velo-pharyngeal activity. Landmarks associated with stop consonants, fricatives, and nasals are acoustically abrupt. Non-abrupt landmarks are associated with semivowels and are regions with gradual spectral changes. A non-abrupt landmark is centered at the local minima in the first formant track which corresponds to the narrowest constriction in the vocal tract during the production of a semivowel. A local maximum in the first formant track corresponding to the widest opening of the vocal tract during the production of a vowel is marked as a vocalic landmark. Non-abrupt and vocalic landmarks together contribute to about 32% of the total number of landmarks (Liu, 1995).

A stop consonant in a typical vowel-consonant-vowel (VCV) context is characterized by three landmarks: stop closure or voicing offset, burst onset, and the voicing onset (Liu, 1996). A decrease in energy in the voicing band (below 400 Hz) due to the closure of the vocal tract is marked as a closure or voicing offset landmark. A burst onset landmark is associated with an abrupt release of frication energy after a closure interval and it may be followed by aspiration. The durations of closure, burst, and voicing onset are dependent on the speaker, speaking style, type of the stop consonant, and the context in which they appear. Typical durations for stop closure and burst are of the order of 50 − 100 ms and 5 − 10 ms, respectively. Voice onset time (VOT) is of the order of 0 – 30 ms and 30 – 150 ms for voiced and unvoiced stops, respectively. Figure 3.1 shows the landmarks associated with voiced and unvoiced stop consonants (/b/ and /p/) in VCV context with vowel /a/. The closure landmark is associated with a gradual fall in energy, particularly in the low frequency range (0 − 400

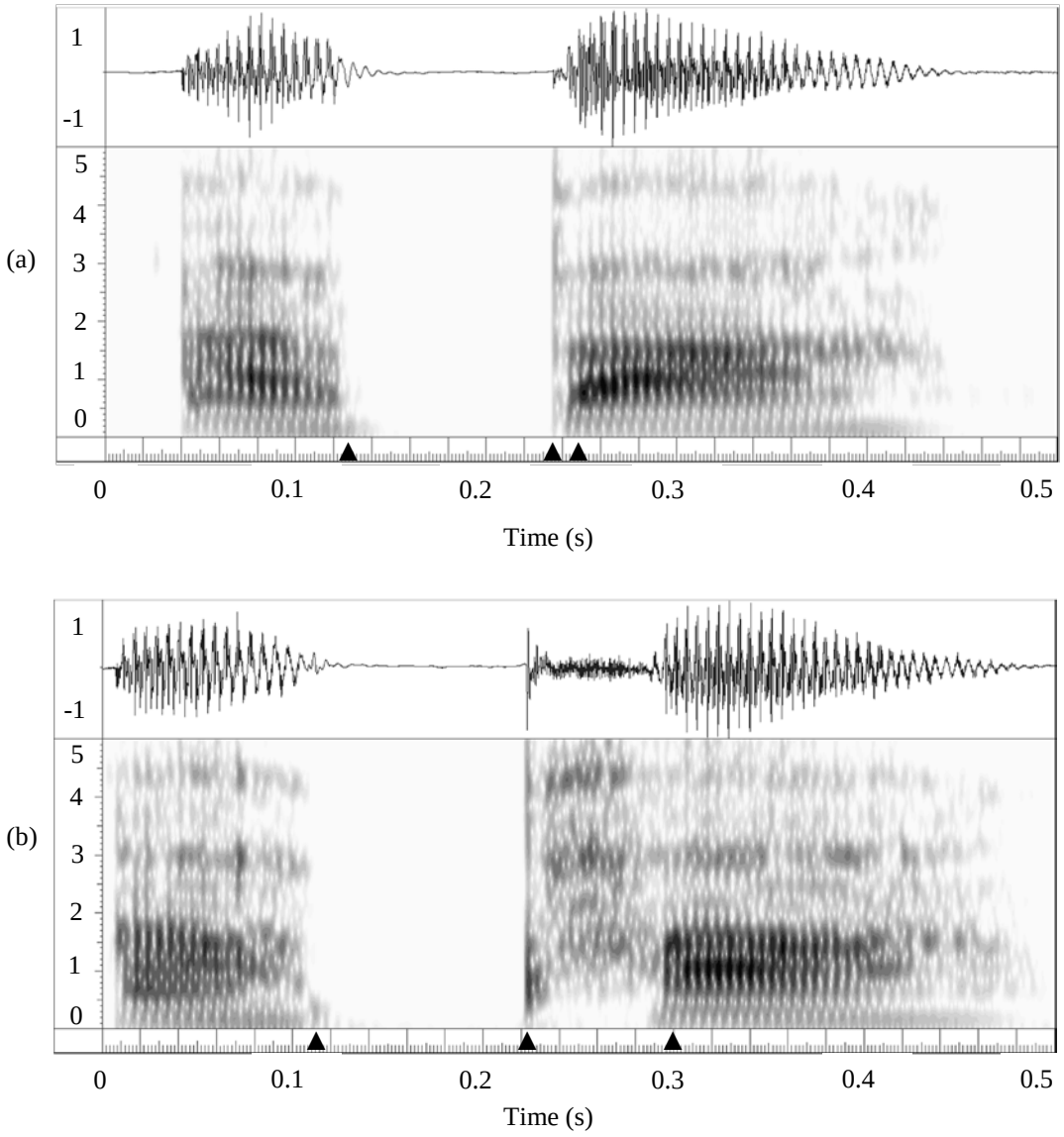


Figure 3.1 Abrupt landmarks in VCV utterances (closure, burst onset, and voicing onset, marked as ▲ along the time axis) with voiced and unvoiced stop consonants: (a) /aba/, (b) /apa/ from a female speaker. Frequency axis of spectrogram in kHz.

Hz). The burst onset landmark is associated with an abrupt rise in energy spread across the entire frequency range, and the voicing onset landmark is associated with a rise in energy mainly in the low frequency range.

A landmark detector for use in a speech intelligibility enhancement application should detect the landmarks with good temporal accuracy with reference to manually located landmarks, preferably with low insertion rates (rates of false detection), and with limited contextual information. Modifications performed around temporally misaligned landmarks (due to poor temporal accuracy of the landmark detector) or at unwanted regions (due to false detections) may adversely affect speech intelligibility. Further, the algorithmic and computational delays involved should be compatible with the requirements of real-time processing. Typical durations of burst and VOT for voiced stops are 5 – 10 ms and 0 – 30 ms,

respectively. For effective enhancement of intensity of burst onsets, we need to detect burst onset landmarks with a temporal accuracy of 5 – 10 ms with respect to the manually marked burst onsets. Temporally misaligned burst onset detection may result in enhancing the intensity of a segment in the closure or in the voicing region, which may not help in improving the perception of stops and may even be detrimental. We need to detect the burst onset with temporal accuracy of 5 – 10 ms to obtain the maximum possible benefit that can be gained out of burst intensity enhancement. Hence a landmark detector with high detection rates at temporal accuracies better than 10 ms is preferred for applications in speech intelligibility enhancement involving segment-specific modifications of short-duration segments. Moderate levels of insertions are tolerable, provided the modification performed around these unwanted regions do not degrade speech intelligibility. A variety of landmark detection techniques have been reported with high detection rates at temporal accuracies of 20 − 30 ms, for applications in speech recognition. But many of such techniques are not capable of time-localizing acoustically abrupt events as their detection rates generally fall to low levels for temporal accuracies of 5 − 10 ms. Ortega et al. (2000) reported poor detection rate and increased insertion rate of the landmark detector as a possible reason for the reduced effectiveness of their automated method for speech modification as compared with the manual method used by Hazan and Simpson (1998).

Liu (1996) proposed a method for landmark detection based on energy variations in 6 spectral bands (0 − 0.4, 0.8 − 1.5, 1.2 − 2.0, 2.0 − 3.5, 3.5 − 5.0, and 5.0 − 8.0 kHz). The processing for landmark detection was performed on speech material sampled at 16 kHz. Magnitude spectrum was computed on Hanning windowed frames of duration 6 ms (96 samples) using 512-point DFT. A high frame rate of 1 frame per ms was used for tracking the acoustically abrupt events. A smoothing was carried out by 20-frame averaging of squared magnitude spectra. Spectral peaks in each of the six bands in the smooth magnitude spectrum (in dB) were considered as indicators of energy variations in the spectral bands and were used as parameters for landmark detection. Rate-of-rise (ROR) functions of these parameters were computed using first difference with time-steps of 26 and 50 ms. These RORs along with segment duration, articulatory, and phonetic class constraints were used to locate and label the landmarks as glottal, sonorant, and burst. The positive peaks (above 9 dB) and negative peaks (below −9 dB) in the ROR of the first band were labeled as voicing onset (+g) and voicing offset (−g) landmarks, respectively. Detection of sonorant onset (+s) and offset (−s) landmarks were attempted within the voicing regions bounded by a pair of +g and −g landmarks. The locations of positive and negative peaks in the RORs of bands 2 − 5 corresponding to frequency range F2 − F4 were grouped and marked as +s and −s landmarks, respectively. The burst onset (b) landmark was characterized by an abrupt rise in the high

frequency energy preceded by a short interval of silence. Location of peaks in the RORs of bands 2 − 5 in the region bounded by a [–g, +g] landmark pair was used for burst onset detection. The silence interval preceding the burst was measured using energy in the bands 3 − 6. Test with 80 sentences (16 speakers × 5 sentences) from TIMIT database (Garofolo et al., 1993) resulted in an overall detection rate of 90%. The detection rates for temporal accuracy of 30 ms were 97, 58, and 93% for glottal, sonorant, and burst landmarks, respectively. The corresponding insertion rates were 3, 23, and 10%. Out of the located landmarks, 88, 85, 68, and 44% were within 30, 20, 10, and 5 ms of the manually located landmarks (Liu, 1995; 1996).

Niyogi and Sondhi (2002) reported an optimal filtering approach for detecting stop consonants in continuous speech using the overall energy, energy above 3 kHz, and Wiener entropy extracted from the short-time spectrum using 5 ms frames, every 1 ms. The optimal filter was designed using data from manually labeled sentences in TIMIT database, to return large values around the stop release bursts and small values otherwise. It performed better than the first difference used by Liu (1996). Use of Wiener entropy along with the energy parameters improved the performance of the landmark detector. Evaluation was performed using 320 sentences (32 speakers × 10 sentences) from TIMIT database. For a temporal accuracy of 20 ms, the detection and insertion rates were 84 and 16%, respectively.

Salomon et al. (2004) used spectral and temporal parameters for landmark detection. The temporal parameters used were related to the envelope (2 − 50 Hz), periodicity (50 − 500 Hz), and fine-structure (above 1 kHz). Envelopes of 60 band-pass channels on the ERB scale were derived using a Hilbert transform based envelope operator. The periodic/aperiodic and onset/offset nature of each channel was obtained and the derivative of log energy was computed with time-steps and window sizes made adaptive to the nature of each channel. A short time-step of 5 ms was used during silence for accurately detecting abrupt onsets. Time-step of 2 pitch periods was used during periodic regions and 30 ms during aperiodic regions. The use of adaptive time-steps and window sizes improved the temporal accuracy of landmark detection. The method detected onsets and offsets of voiced, sonorant, and obstruent sounds in continuous speech. In a test using 120 sentences (8 female, 16 male speakers × 5 sentence) from TIMIT database, the overall detection rate of the method for a temporal accuracy of 50 ms was 80.2% with insertion rate of 8.7%. Compared with other landmarks, higher detection rates were obtained for stop closures (90.9%) and release bursts (86%). The method detected nearly 99% of unvoiced stop release bursts. Ortega et al. (2000) used a broad-class HMM classifier for detecting regions for modification in an automated method for speech intelligibility enhancement. The detection rates of the method for bursts,

fricatives, nasals, vowel onsets, and offsets were 80, 75, 90, 78, and 76% respectively. The corresponding insertion rates were 25, 9, 24, 16, and 17%, respectively.

For practical application of the speech intelligibility enhancement based on speech characteristic modification, the landmark detection should involve small algorithmic and computational delays. Investigations by Stone and Moore (2005) have shown that processing delays of the order of 14 to 30 ms are tolerable for listeners with hearing loss and higher delays may result in audio-visual asynchrony and disturbance in speech perception. In an approach for automated enhancement of speech intelligibility, Colotte and Laprie (2000) used energy in the spectral bands $0.6 - 1.0$ kHz and $3.6 - 6$ kHz along with an MFCC based spectral variation function to detect the stop and fricative landmarks. The speech modification involved amplification of stop and fricative landmarks and time-scale expansion (by 1.8 to 2) of short duration segments centered at the local maxima in the spectral variation function. The detection rate was 82% for a temporal accuracy of 20 ms. Insertions were not considered a serious problem because the processing of such landmarks was not detrimental to speech intelligibility.

Skowronski and Harris (2006) used a voiced/unvoiced detector based on spectral flatness measure (SFM) in a speech intelligibility enhancement technique involving redistribution of energy between voiced and unvoiced segments. The measure was calculated as the ratio of geometric mean to arithmetic mean of the magnitude spectrum, for 20 ms windows with 50% overlap. The value of SFM approached unity during unvoiced segments and was small during voiced segments. A Schmitt trigger based boundary decision using two thresholds (0.36 and 0.47) was used to demarcate voiced and unvoiced regions.

The following sections describe the landmark detection techniques investigated for modification of speech characteristics using the acoustic properties of clear speech for intelligibility enhancement. Focus is on the detection of landmarks associated with stop consonants (/b, d, g, p, t, k/). The weak and transient nature of stop consonants poses difficulty in their perception by hearing impaired listeners as well as by normal-hearing listeners in noisy environments. Speech perception in noise can be improved by providing listeners enhanced access to the acoustic landmarks (Li and Loizou, 2008). The objective of the investigation is to derive an effective set of parameters, time-step, and distance measure for real-time detection of stop consonant landmarks in continuous speech. Importance is given to improve the temporal accuracy of landmark detection and to reduce the computational requirements. Four landmark detection methods based on (i) subband energy and centroid (EC), (ii) parameters of Gaussian mixture model (GMM) of the spectrum, (iii) spectral moments (SM), and (iv) spectral moments with a tone added to the signal (SMTA) are investigated. The performance of the landmark detection method is evaluated using manually

labeled VCV utterances and sentences from TIMIT database. The results from these investigations are used to select a method suited for real-time implementation.

3.2 Subband energy and centroid based method (EC)

This investigation is based on the observation that centroid frequency within a spectral band contains important information regarding the distribution of energy and can be used as a parameter for landmark detection. Spectral subband centroids have formant-like features and can be easily estimated from the spectrum. It has been reported (Paliwal, 1998) that use of subband centroids as supplementary features to cepstral features improved speech recognition accuracy by $4-6\%$.

In our method, variation of parameters related to energy and centroid in spectral subbands were combined to locate the abrupt spectral transitions associated with the landmarks. The rate-of-change (ROC) functions of band energies and centroids were multiplied together and the products were summed across the bands to get a single parameter called transition index, indicative of the overall spectral variation.

3.2.1 Computation of subband energy, centroids, and transition index

For speech signal sampled at 10 kHz, short-time magnitude spectrum was computed using 512-point DFT of 6 ms Hanning windowed frames, taken every 1 ms. Use of short analysis window suppresses the harmonic structure in the spectrum, and the high frame rate helps in precisely locating the landmarks. The spectrum was divided into five non-overlapping bands: $0-0.4$, $0.4-1.2$, $1.2-2.0$, $2.0-3.5$, $3.5-5.0$ kHz. Parameters from band 1 tracked the glottal vibrations and those from bands $2-5$ were related to variations in the frequency range of F1 − F4. The magnitude spectrum was smoothed by a 20-frame moving average to get $|X(n,k)|$. The spectral peak (in dB) of band *b* for frame *n* was calculated as

$$E_b(n) = 20\log_{10}(\max(|X(n,k)|, k_1 \le k \le k_2)) \tag{3.1}$$

where $k_1$and k_2 are the lower and upper indices for band b. The centroid of band *b* for frame *n* was calculated as

$$f_b(n) = (f_s/N)\sum_{k_1}^{k_2} k|X(n,k)| / \sum_{k_1}^{k_2}|X(n,k)| \tag{3.2}$$

where *N* is the number of points in the DFT computation and f_s is the sampling frequency. A rate-of-change (ROC) function based on first difference with a fixed time-step was used to measure the variation of parameters. ROCs of E_b and f_b were computed every 1 ms as

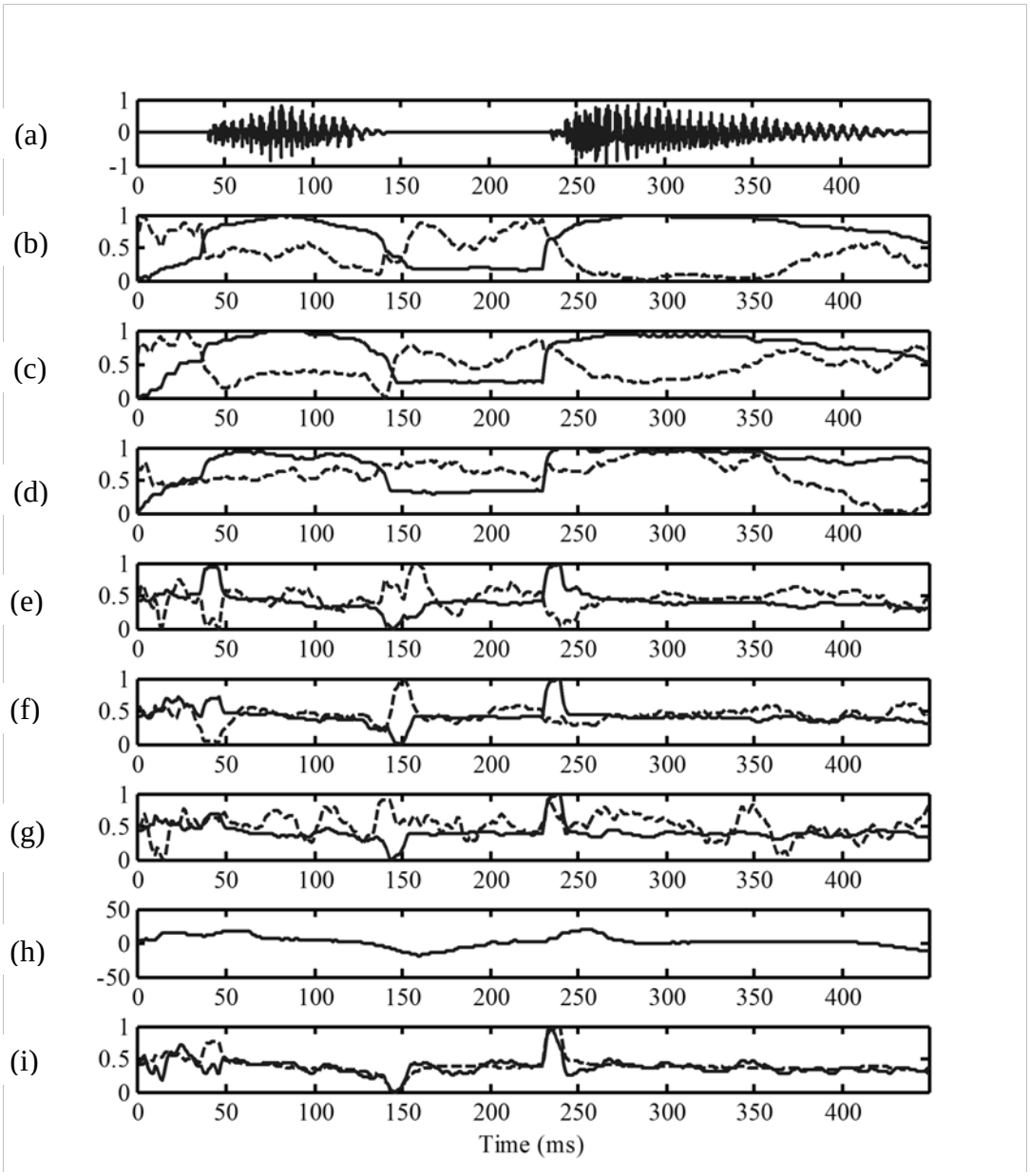


Figure 3.2 Landmark detection method EC for /aba/: (a) signal waveform, (b) band 3 energy (solid) and centroid (dotted), (c) band 4 energy (solid) and centroid (dotted), (d) band 5 energy (solid) and centroid (dotted), (e) band 3 ROC functions for band energy (solid) and centroid (dotted) (f), band 4 ROC functions, (g) band 5 ROC functions, (h) ROC for band 1 energy, and (i) transition indices $T_{rEC}(n)$ (solid) and $T_{rE}(n)$ (dotted).

$$E_b'(n) = E_b(n) - E_b(n-K) \tag{3.3}$$

$$f_b'(n) = f_b(n) - f_b(n-K) \tag{3.4}$$

where *K* is the time-step. The ROCs remained near-to-zero during steady-state segments and had well defined peaks during abrupt spectral transitions.

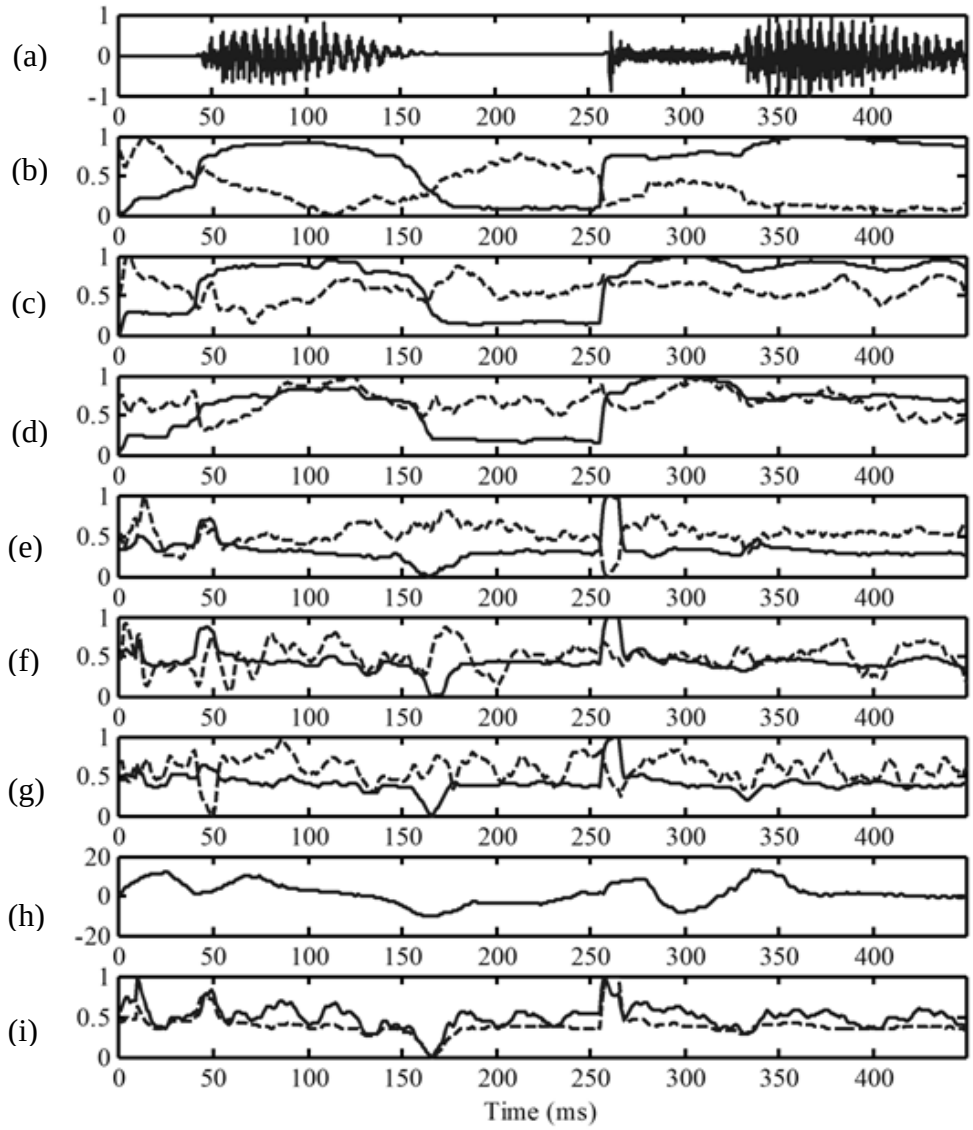


Figure 3.3 Landmark detection method EC for /apa/: (a) signal waveform, (b) band 3 energy (solid) and centroid (dotted), (c) band 4 energy (solid) and centroid (dotted), (d) band 5 energy (solid) and centroid (dotted), (e) band 3 ROC functions for band energy (solid) and centroid (dotted), (f) band 4 ROC functions, (g) band 5 ROC functions, (h) ROC for band 1 energy, and (i) transition indices $T_{rEC}(n)$ (solid) and $T_{rE}(n)$ (dotted).

The detection of voicing offset (−g) and voicing onset (+g) landmarks was performed using ROC of E_b in band 1 computed using (3.3) with a time-step *K* of 26 ms. The location of peak in the ROC with value above 9 dB was taken as the voicing onset (+g) landmark. Location of negative peak with value below −9 dB was taken as the voicing offset (−g) landmark.

To locate the simultaneous variation of energy and its frequency distribution in each spectral band, point-by-point product of the normalized band energy and centroid ROCs were

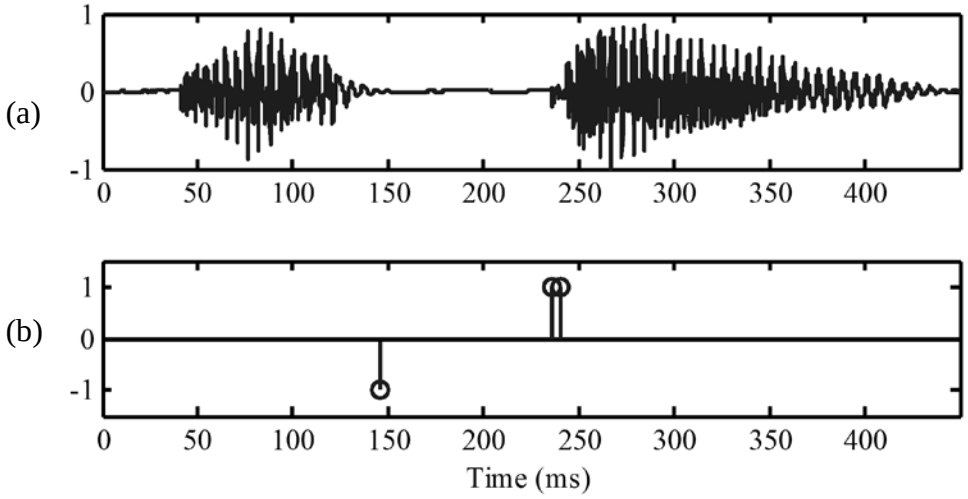


Figure 3.4 Landmark detection method EC for /aba/: (a) signal waveform, (b) voicing offset (−g), burst onset (b) and voicing onset (+g) landmarks.

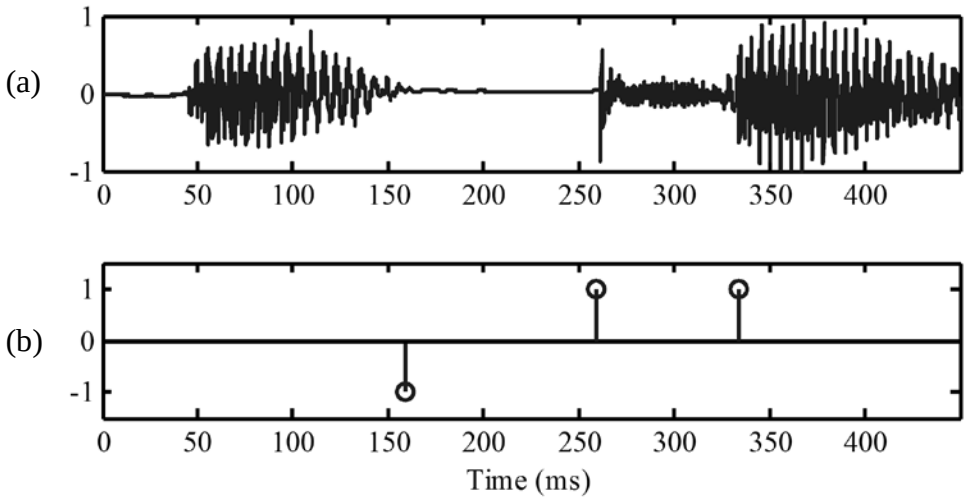


Figure 3.5 Landmark detection method EC for /apa/: (a) signal waveform, (b) voicing offset (−g), burst onset (b), and voicing onset (+g) landmarks.

taken along the frame index *n.* A measure of the overall spectral variation in the frequency range of interest was obtained by summing these products across the spectral bands. For detection of stop release bursts and frication onsets, ROCs from bands 3 − 5 were used for computation of a transition index. Band 1 and band 2 were not included in the computation to eliminate the effect of voice bar that may occur for voiced stop consonants.

The ROCs were computed using (3.3) and (3.4) with a time-step *K* of 10 ms. These ROCs were normalized to the range 0 to 1 by offset and scaling. The transition index based on band energies and centroids $T_{rEC}(n)$, computed as

$$T_{rEC}(n) = \sum_{b=3}^{5} |E'_b(n) f'_b(n)| \tag{3.5}$$

was found to be steady during vowel segments and had prominent peaks during abrupt spectral transitions. A transition index $T_{rE}(n)$ based only on band energies was also computed as

$$T_{rE}(n) = \sum_{b=3}^{5} |E'_b(n)| \tag{3.6}$$

to compare its performance with that of $T_{rEC}(n)$. Both $T_{rE}(n)$ and $T_{rEC}(n)$ were normalized to the range of 0 to 1. Spectral transitions associated with release bursts of stop consonants and onsets of frication with abrupt rise of energy in the high frequency bands were marked with peaks with values close to 1 in the transition index contour. The transition index approached zero during stop closures and frication offsets.

As an example of processing, the signal waveform, plots of band energies and centroids in the bands 3 to 5, corresponding ROCs, ROC of band-1 energy used for detection of voicing offset and onset, and the transition index contours $T_{rE}(n)$ and $T_{rEC}(n)$ used for burst onset detection are shown in Figure 3.2 for VCV utterance /aba/. The corresponding plots for /apa/ are shown in Figure 3.3. In a VCV utterance, burst onset (b) is located in the region bounded by voicing offset (−g) and voicing onset (+g) landmarks. The location of the most prominent peak in the transition index contour between −g and +g landmarks was taken as the burst onset landmark. Figures 3.4 and 3.5 illustrate the detection of the landmarks associated with the VCV utterance /aba/ and /apa/, respectively.

3.2.2 Evaluation of the EC method

The method was evaluated using 180 VCV utterances involving six stop consonants (/b, d, g, p, t, k/) and 3 vowels (/a, i, u/) recorded from 5 male and 5 female speakers. The location of −g, b, and +g landmarks in the utterances were manually marked. The detection rates at different temporal accuracies for voicing offsets and onsets are listed in Table 3.1. Table 3.2 lists the detection rates for burst onsets using T_{rE} and T_{rEC}. The burst onset detection rates are comparable for the two methods at moderate temporal accuracies. Improved performance is obtained for T_{rEC} for temporal accuracies of 5 and 7 ms, indicating the usefulness of band centroids in localizing the burst onset landmarks.

The method was also evaluated using 50 manually annotated conversational style sentences (3 female and 2 male speakers × 10 sentences) from TIMIT database. Based on examination of ROCs for band 1 for a number of sentences in the test material, detection of −g and +g landmarks were performed with threshold levels of −6 and +6 dB, respectively. In addition to the regions bounded by [–g, +g] landmark pair, the search for burst onset was extended also around isolated –g and +g landmarks to the regions [–g, –g + 50] and [+g − 50, +g], respectively. The locations of peaks in the transition index contour $T_{rEC}(n)$ in the search

Table 3.1 Landmark detection method EC: Detection rates (%) for voicing offsets (−g) and onsets (+g) in VCV utterances.

Landmarks	Temporal accuracy (ms)					
(no. of tokens)	≤30	≤20	≤15	≤10	≤7	≤5
−g (180)	93	83	73	64	56	36
+g (180)	99	98	98	93	83	76

Table 3.2 Landmark detection method EC: Detection rates (%) for burst onsets using $T_{rE}(n)$ and $T_{rEC}(n)$ in VCV utterances.

Transition index	Temporal accuracy (ms)					
	≤30	≤20	≤15	≤10	≤7	≤5
$T_{rE}(n)$	93	91	91	89	62	21
$T_{rEC}(n)$	93	90	88	86	66	44

Table 3.3 Landmark detection method EC: Detection rates (%) for stop consonant landmarks in TIMIT sentences.

Landmark	Temporal accuracy (ms)					
(no. of tokens)	≤30	≤20	≤15	≤10	≤7	≤5
−g (270)	90	80	63	40	27	19
+g (232)	96	91	82	71	60	45
b (306)	92	92	75	60	45	33

intervals, with values above an empirically set threshold were matched with the manual transcription of the sentences. The threshold was set based on observation of the values of the transition index at valid burst onsets in a number of utterances. The detection rate of the method is listed in Table 3.3, with the number of tokens in the class given in brackets. In this evaluation, the TIMIT labels corresponding to release bursts (/b, d, g, p, t, k/) of stop consonants were grouped as stops. The detection rates for stop release bursts were 92% for temporal accuracy of 20 ms and 60% for temporal accuracy of 10 ms.

The method based on band energy and centroids uses parameters that can be easily estimated from the spectrum. The method has good detection rates at moderate levels of temporal accuracy (20 − 30 ms). The centroids estimated from spectral bands improved the detection rates at temporal accuracy of 7 and 5 ms. Compared with band energy parameters, centroids had more variations during steady-state segments and closure intervals as seen in Figure 3.2 and Figure 3.3. The variation during steady-state segments could be because of the use of fixed band boundaries. Relaxing this constraint is expected to improve the performance of centroid parameter for landmark detection. The variations during closure intervals seem to

be because of the undefined nature of centroid computed using (3.2) during segments with very low energy. Alternate methods for estimation of parameters, which address the problems related to the band boundary constraints and undefined nature of centroid during closure intervals, are expected to improve the performance of the landmark detector.

3.3 Gaussian mixture model (GMM) based method

Energy variations in spectral bands with fixed boundaries capture the vocal tract resonances in an approximate way, and may not be able to model speaker related spectral variability. Liu (1995) reported nearly 7% variation in detection rates due to speaker dependent variability of the parameters used for landmark detection. In case of multiple spectral prominences in a single band or spectral prominences spread across bands, band energy parameters may not meaningfully represent the spectral variations.

Our investigation is based on the assumption that the use of parameters derived from a spectral modeling approach which can adapt to the dynamic nature of the spectrum may improve the detection rate and temporal accuracy of landmark detection. A Gaussian mixture model (GMM) of the short-time speech spectrum provides a parametric representation of the spectral envelope using a weighted sum of Gaussian functions. An approximation with a small error can be obtained, for all classes of sounds using a small number of Gaussian components in the mixture model. Zolfaghari and Robinson (1996) used a GMM based parametric scheme for extracting formant-like features. GMM parameters have been used for improving speech recognition in noisy environments and for performing spectral modifications (Stuttle and Gales, 2002; Stuttle, 2003; Zolfaghari et al., 2006; Nguyen and Akagi, 2009).

3.3.1 Estimation of Gaussian parameters

The short-time log magnitude spectrum was modeled by a mixture of Gaussian functions. Log-magnitude has been used in the literature as it provides dynamic range compression for supra-threshold values. A preliminary investigation showed that the 4-component GMM followed the envelope of log-magnitude spectrum much better than that of magnitude and squared magnitude spectra. Further, use of first difference operation on the parameters in the log-scale gives an advantage of working with relative values, eliminating the need for gain normalization across utterances. The log magnitude spectrum was computed, for speech signal sampled at 10 kHz, using 512-point DFT on 6 ms Hanning windowed frames, taken every 1 ms. Use of a short duration window suppressed the pitch harmonics in the spectrum and thus prevented the Gaussian components from tracking non-formant peaks (Zolfaghari and Robinson, 1996; Nguyen and Akagi, 2009). The high frame rate helped in accurately tracking the fast spectral variations. The magnitude spectrum was smoothed by a 50-point

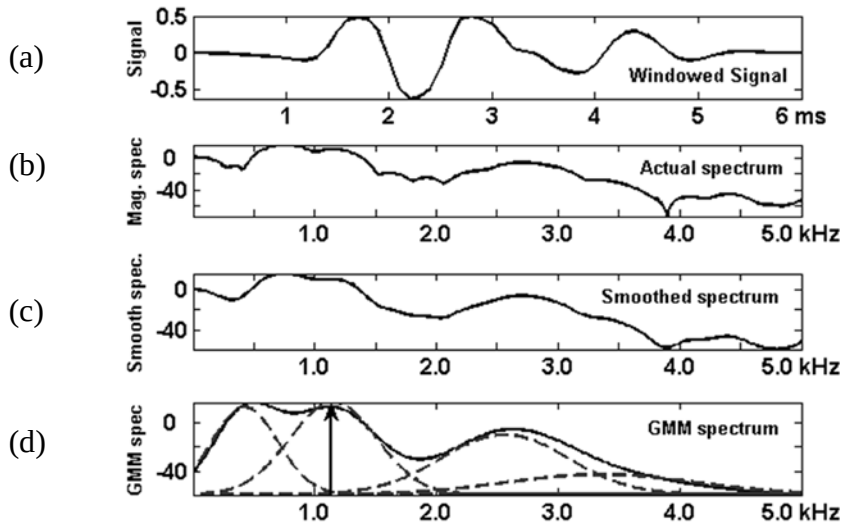


Figure 3.6 Fitting GMM on spectrum of vowel /*a*/: (a) windowed segment of 6 ms, (b) log magnitude spectrum, (c) smoothened spectrum, (d) GMM approximated spectrum with dotted lines indicating the individual Gaussian components.

median filter, taken along the frequency index *k*. The smoothed log magnitude spectrum $S_n(k)$ was approximated by a weighted sum of *M* Gaussian functions as

$$\hat{S}_n(k) = \sum_{g=1}^{M} w_{gn} G(\mu_{gn}, \sigma_{gn}, k) \tag{3.7}$$

where w_{gn}, μ_{gn} and σ_{gn}^2 represent the weight, mean, and variance, respectively of the g^{th} Gaussian in the mixture model for frame *n*. A reasonably good approximation of the speech spectrum is possible with 4 or 5 Gaussian components in the mixture model (Zolfaghari et al., 2006; Jayan and Pandey, 2008; 2009). As the approximation errors for the two do not significantly differ, we have used the 4-component model.

The GMM parameters were estimated using expectation maximization (EM) algorithm (Stuttle, 2003; Duda et al., 2004). With a given initialization, it iteratively computed the maximum likelihood estimates of the model parameters. The spectrum $S_n(k)$ was viewed as a histogram with rectangular bins placed at each frequency index *k*. Iterations were started with an initial set of parameters. The probability $p(g \mid k)$ that frequency measurement *k* came from the Gaussian component *g*, was evaluated as

$$p(g|k) = w_{gn} G(\mu_{gn}, \sigma_{gn}, k) / \sum_{g=1}^{M} w_{gn} G(\mu_{gn}, \sigma_{gn}, k) \tag{3.8}$$

The new mixture weights, means, and variances were calculated as

$$\hat{w}_{gn} = \sum_{k=1}^{N/2} S_n(k) p(g|k) / \sum_{k=1}^{N/2} S_n(k) \tag{3.9}$$

$$\hat{\mu}_{gn} = \sum_{k=1}^{N/2} kS_n(k)p(g|k) / \sum_{k=1}^{N/2} S_n(k)\,p(g|k) \tag{3.10}$$

$$\hat{\sigma}_{gn}^2 = \sum_{k=1}^{N/2} \left(k - \mu_{gn}\right)^2\ S_n(k)p(g|k) / \sum_{k=1}^{N/2} S_n(k)\,p(g|k) \tag{3.11}$$

where N is the number of points in the DFT computation. The parameters were used in the next iteration. The iterations were continued until the changes in the parameter values in successive iterations became less than a set threshold or the number of iterations reached a set limit.

The choice of initial parameters for the EM algorithm affects the number of iterations needed and the solutions obtained. In our implementation, the mixture weights were initialized with equal values. The means and the variances were initialized with the average values of vowel formant frequencies and extreme values of formant bandwidths, respectively, of the corresponding formants (first: 600, 160; second: 1200, 200; third: 2400, 300; fourth: 3600, 400 Hz, as used by Deller et al., (2000)). The 3-dB bandwidth B_g of a Gaussian distribution is related to its standard deviation σ_g by

$$B_g = 2.35\sigma_g \tag{3.12}$$

For sampling frequency f_s and N-point DFT, the initialization values for the g^{th} component are obtained from the values of the average formant F_g and extreme formant bandwidth B_g by using the correspondence

$$\mu_g = NF_g/f_s \tag{3.13}$$

$$\sigma_g = NB_g/(2.35f_s) \tag{3.14}$$

These initializations were found to track the spectral changes in the speech signals from a number of male and female speakers. The maximum number of iterations was set at 12 because no significant decrease in approximation error was observed by increasing the number of iterations. Figure 3.6 shows modeling for a 6 ms segment of vowel /a/ spoken by a male speaker, with the peaks in the GMM approximation generally matching the resonance peaks in the spectrum.

3.3.2 Detection of burst onset landmarks

A rate of change function defined on the GMM parameters along with a voicing onset-offset detector and a spectral flatness measure (Skowronski and Harris, 2006) was used for the detection of stop consonant landmarks. The amplitudes of the Gaussian modeled spectral envelope at the four mean locations (A_g) were found to be more consistently related to the spectral changes, compared with the mixture weights. The parameters used for landmark detection included the means (μ_g), square root of the variances (σ_g), and amplitude of the

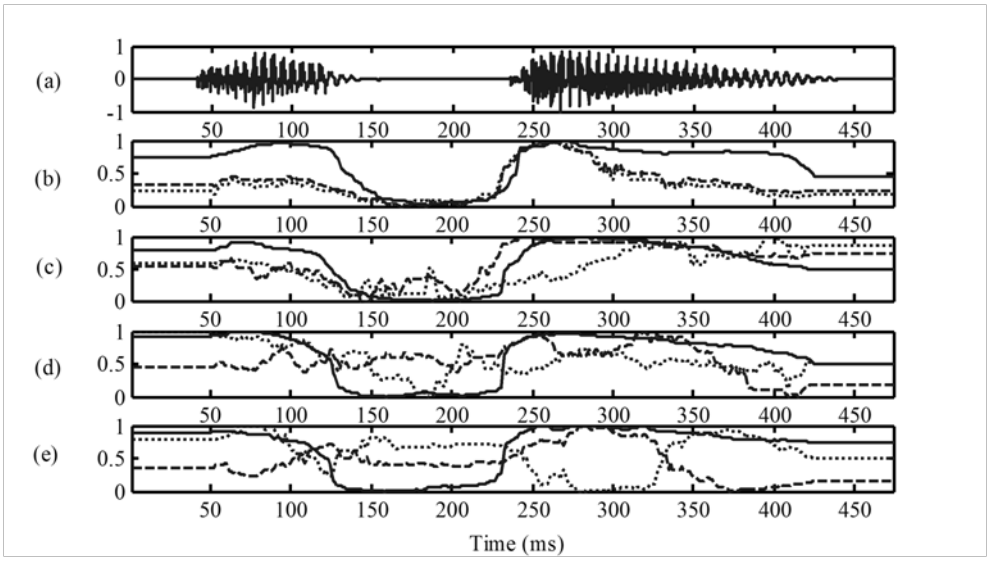


Figure 3.7 Landmark detection method GMM for /aba/: (a) signal waveform, (b) Gaussian 1 (*A*: thick, μ: dashed, σ: dotted), (c) Gaussian 2, (d) Gaussian 3, (e) Gaussian 4.

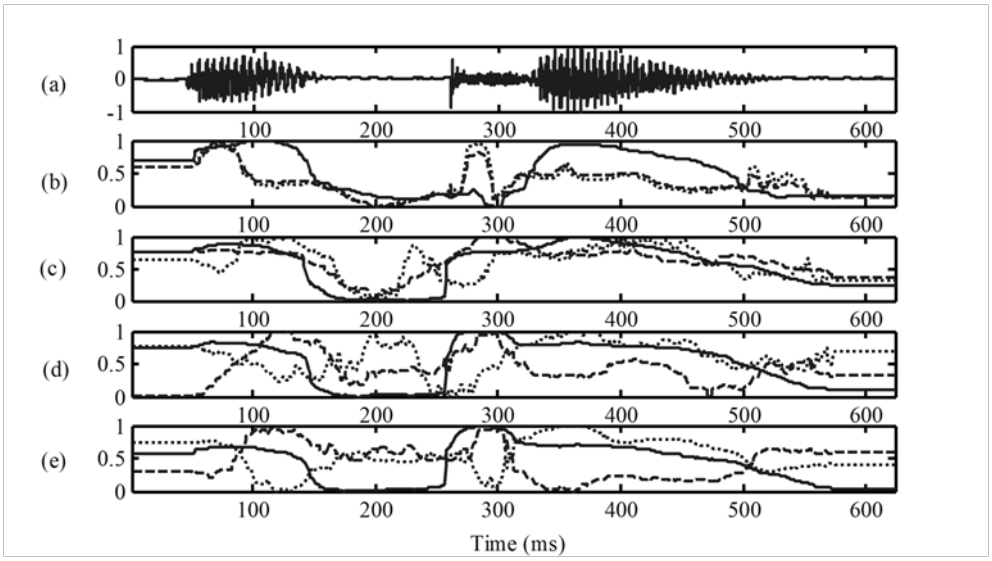


Figure 3.8 Landmark detection method GMM for /apa/: (a) signal waveform, (b) Gaussian 1 (*A*: thick, μ: dashed, σ: dotted), (c) Gaussian 2, (d) Gaussian 3, (e) Gaussian 4.

Gaussian modeled envelope at the mean locations (A_g) of the four Gaussian components. A 30-point median filtering was applied on the parameter tracks. It helped in smoothening them during steady-state segments without significantly smearing the variations corresponding to abrupt spectral transitions. The smoothed parameters denoted as A'_g, μ'_g, σ'_g were used for calculating a rate-of-change (ROC) function given as

$$r_c(n) = r_A(n) r_\mu(n) r_\sigma(n) / R \tag{3.15}$$

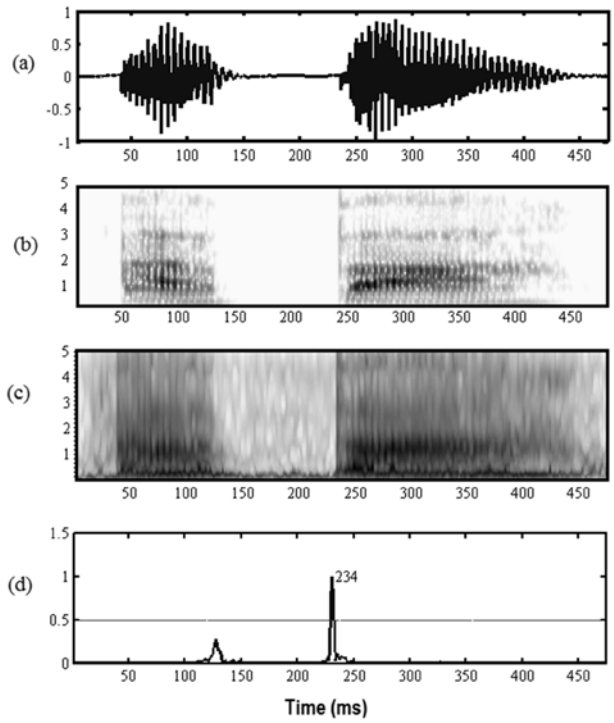


Figure 3.9 Landmark detection method GMM for /aba/: (a) signal waveform, (b) spectrogram (frequency in kHz), (c) GMM spectrogram (frequency in kHz), (d) GMM ROC.

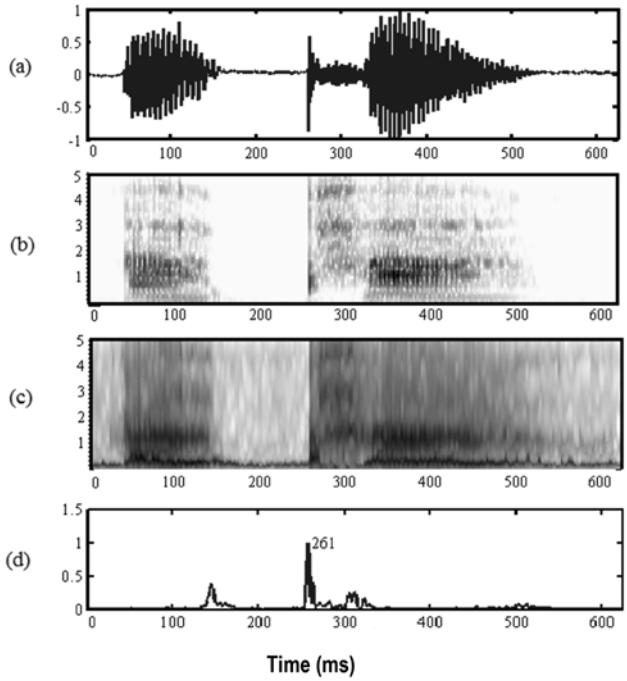


Figure 3.10 Landmark detection method GMM for /apa/: (a) signal waveform, (b) spectrogram (frequency in kHz), (c) GMM spectrogram (frequency in kHz), (d) GMM ROC.

where

$$r_A(n) = \sum_{g=1}^{4} \left| A'_g(n) - A'_g(n-K) \right| \tag{3.16}$$

$$r_\mu(n) = \sum_{g=1}^{4} \left| \mu'_g(n) - \mu'_g(n-K) \right| \tag{3.17}$$

$$r_\sigma(n) = \sum_{g=1}^{4} \left| \sigma'_g(n) - \sigma'_g(n-K) \right| \tag{3.18}$$

and *R* is used to scale the maximum value of r_c to 1. A short time-step *K* of 2 ms helped to suppress relatively slow spectral variations associated with semivowels, voicing offsets, etc. The product operation on individual ROCs ensured the strong peaks to occur at the points of sharp spectral variations. These peaks in the $r_c(n)$ contour indicated the possible location of release burst onsets. Voicing offsets (−g) and onsets (+g) were located by the method described earlier in Section 3.2.1. A measure of spectral flatness (SFM) was calculated as the ratio of geometric mean to the arithmetic mean of the magnitude spectrum, for 20 ms Hanning windowed frames, every 1 ms as

$$\mathrm{SFM}(n) = \left(\prod_{k=1}^{N/2} |X_n(k)| \right)^{2/N} \Big/ (2/N) \left(\sum_{k=1}^{N/2} |X_n(k)| \right) \tag{3.19}$$

It is low for voiced frames with peaky spectra and close to 1 for frication with flat spectra (Skowronski and Harris, 2006). For sentence material, burst onset detection was attempted around voicing onsets paired with preceding voicing offsets, isolated voicing onsets, and isolated voicing offsets. The burst onset candidates were validated by ensuring a closure for a minimum duration of 10 ms preceding it and a rise in the SFM contour above 0.5 in the vicinity of the burst onset candidate. As examples of processing, Figure 3.7 and Figure 3.8 show the median smoothened Gaussian parameter tracks for /aba/ and /apa/, respectively. For /aba/, the Gaussian modeled spectrograms and GMM ROCs used for burst onset detection are shown in Figure 3.9. Similar plots for /apa/ are shown in Figure 3.10.

3.3.3 Evaluation of the GMM method

The method was evaluated using the speech material consisting of VCV utterances and sentences as used for evaluation of the EC method and described earlier in Section 3.2.2. Stop landmark detection rates in VCV utterances for different temporal accuracies are given in Table 3.4. The burst detection rates were 98, 96, 93, 92, 91, and 90% at temporal accuracies of 30, 20, 15, 10, 7, and 5 ms respectively. At all temporal accuracies, the detection rates using the GMM method were much better than those obtained using the EC method. Out of the release bursts, 90% got detected within a temporal accuracy of 5 ms, compared to 44% obtained using the EC method. In the evaluation using a set of 50 conversational style sentences from the TIMIT database, the TIMIT labels corresponding to the release bursts of stop consonants (/b, d, g, p, t, k/) were grouped as stops. Instead of comparing the ROC with an empirical threshold as in the earlier method EC, the search was restricted to stop release bursts in [−g, +g] regions. In case of unpaired −g and +g, the search interval was limited to [−g, −g+50] and [+g−50, +g], respectively. The most prominent peak in GMM ROC in the search interval was taken as the stop release burst. This resulted in some insertion errors in the

Table 3.4 Landmark detection method GMM: Detection rates (%) for stop consonant landmarks in VCV utterances.

Landmarks	Temporal accuracy (ms)					
(no. of tokens)	≤30	≤20	≤15	≤10	≤7	≤5
−g (180)	93	83	73	64	56	36
+g (180)	99	98	98	93	83	76
b (180)	98	96	93	92	91	90

Table 3.5 Landmark detection method GMM: Detection rates (%) for stop consonant landmarks in TIMIT sentences.

Landmarks	Temporal accuracy (ms)					
(no. of tokens)	≤30	≤20	≤15	≤10	≤7	≤5
−g (270)	90	80	63	40	27	19
+g (232)	96	91	82	71	60	45
b (306)	98	97	95	90	80	73

Table 3.6 Landmark detection method GMM: Insertions rates for TIMIT sentences.

Type of transition	Insertion rate (%)
Clicks, glottal stops	8
Vowel-semivowel	4
Stop to /l/, /r/	1

sentence material. Performance of the detection process for different types of landmarks is summarized in Table 3.5. Out of the total 306 stops, the method was able to detect 223 stops (~73%) within 5 ms of the TIMIT transcriptions. Closure onsets were evaluated on 270 tokens (marked as ‘bcl’, ‘dcl’, ‘gcl’, ‘pcl’, ‘tcl’, ‘kcl’ in TIMIT transcription) having preceding voiced segments. Voicing onset detection was evaluated on 232 stop release bursts followed by voiced segments. There were a total of 39 (~13%) insertions, which are described by the phoneme transitions listed in Table 3.6. Affricate detections were not counted as insertions. The insertions were mainly due to burst like clicks preceded by low energy segments and abrupt spectral transitions caused by glottal stops marked as ‘q’ in the TIMIT transcription. Using GMM parameters, it was possible to detect over 90% of stop release bursts in VCV syllables and 73% of stop bursts in TIMIT sentences within 5 ms of the manually annotated landmarks.

The use of parameters from the Gaussian mixture model improved the detection rates at temporal accuracy below 10 ms for both VCV utterances and sentence material. For a

temporal accuracy of 10 ms, the detection rates for bursts using the EC method were 86% and 60% for the VCV utterances and TIMIT sentences, respectively. The corresponding values with the GMM method were 92% and 90%.

The detection of burst onset landmarks using this method involves scanning for closures bounded by [−g, +g] pair, unpaired −g, and unpaired +g. Therefore, the method has an algorithmic delay of nearly 400 ms. Iterative calculations and scanning operations involved in the method make it highly computation intensive. Its Matlab based implementation running on a 1.4 GHz Intel Pentium-M processor based PC took 2 − 3 minutes for detection of landmarks in a typical TIMIT sentence of duration 4 − 5 s. Because of the excessive algorithmic delay and computation requirement associated with it, the method is not suited for real-time implementation.

3.4 Method based on spectral moments (SM)

Spectral moments are related to the spectral shape and have been used as supplementary parameters for classification of Mandarin stops (Lin and Wang, 2008). In this section, we present investigations on the use of spectral moments for landmark detection. A rate of change measure based on Mahalanobis distance (Mahalanobis, 1936; Deller et al., 2000) to combine variations of a set of parameters with different dynamic ranges and correlations and to get a single measure indicative of the overall spectral variation is presented. The effect of time-steps on the detection rate is investigated. The effectiveness of band energies and spectral moments as parameters for landmark detection, individually and in a combined fashion has also been investigated.

3.4.1 Computation of spectral moments

For speech sampled at 10 kHz, 512-point DFT was computed for 6 ms Hanning windowed frames, every 1 ms. The magnitude spectrum was smoothed by a 20-frame moving average. The band energy parameters $E_b(n)$ in frequency bands 1.2 − 2.0, 2.0 − 3.5, and 3.5 − 5.0 kHz and denoted as $E_{b1},\ E_{b2},\ E_{b3}$, respectively, were computed using (3.1). The first four spectral moments (i) centroid, (ii) standard deviation, (iii) skewness, and (iv) kurtosis indicate the frequency of concentration of the spectral energy, spread of energy around this location, the symmetry of the spectrum, and its peakiness, respectively. For computing the spectral moments, the smoothed spectrum $X(n,k)$ was normalized as

$$p(n,k) = |X(n,k)| / \sum_{k=1}^{N/2} |X(n,k)| \tag{3.20}$$

where N is the number of points in the DFT computation. The centroid of the spectrum was computed as

$$F_c(n) = \sum_{k=1}^{N/2} f_k p(n,k) \tag{3.21}$$

where f_k is the frequency in Hz corresponding to the DFT bin with index *k*. The second, third, and fourth moments related to the variance $F_\sigma(n)$, skewness $F_s(n)$, and kurtosis $F_k(n)$, respectively, were computed as

$$F_m(n) = \left[\sum_{k=1}^{N/2} \left(f_k - F_c(n)\right)^m p(n,k)\right]^{1/m} \tag{3.22}$$

where m = 2 for $F_\sigma(n)$, m = 3 for $F_s(n)$, and m = 4 for $F_k(n)$.

3.4.2 Computation of rate of change

Landmark detection involves locating regions in the speech signal with a significant variation in a set of parameters characterizing the landmark. The parameters used for landmark detection may have different magnitude scales and they may be correlated to a certain extent. Mahalanobis distance (Mahalanobis, 1936; Deller et al., 2000) can be used to take care of the scale differences and the correlations of the parameters and to get a single rate of change indicating the overall variation of parameters. Mahalanobis distance between point $\mathbf{y_1}$ and point $\mathbf{y_2}$ is defined as

$$d = ((\mathbf{y_1} - \mathbf{y_2}) \sum^{-1} (\mathbf{y_1} - \mathbf{y_2})^T)^{0.5} \tag{3.23}$$

where $\sum$ is the covariance matrix of the vector containing $\mathbf{y_1}$ and $\mathbf{y_2}$. Three methods were investigated for the computation of the covariance matrix using parameters from (i) 20 frames preceding the current frame, (ii) the entire utterance, and (iii) portion of the utterance excluding silence and very low energy segments. Examination of individual values in the covariance matrix and observation of the rate of change (ROC) tracks from these three approaches indicated the third method to be more consistent (Jayan et al., 2011). The threshold used to demarcate speech and silence was kept 20 dB below the maximum signal level in the utterance. The rate of change function denoted as ROC-MD was computed as

$$\mathrm{ROC_{MD}}(n) = (\mathbf{d}(n) \sum^{-1} \mathbf{d}(n)^T)^{0.5} \tag{3.24}$$

where

$$\mathbf{d}(n) = \mathbf{y}(n) - \mathbf{y}(n-K) \tag{3.25}$$

and the covariance matrix $\sum$ is precomputed from the selected parameter set, $\mathbf{y}(n)$ is the parameter set of the current frame *n*, and *K* is the time step. As examples of processing, Figure 3.11 and 3.12 show the waveform, energy tracks, spectral moments, and ROC-MD using normalized vertical scales for utterances /aba/ and /apa/, respectively.

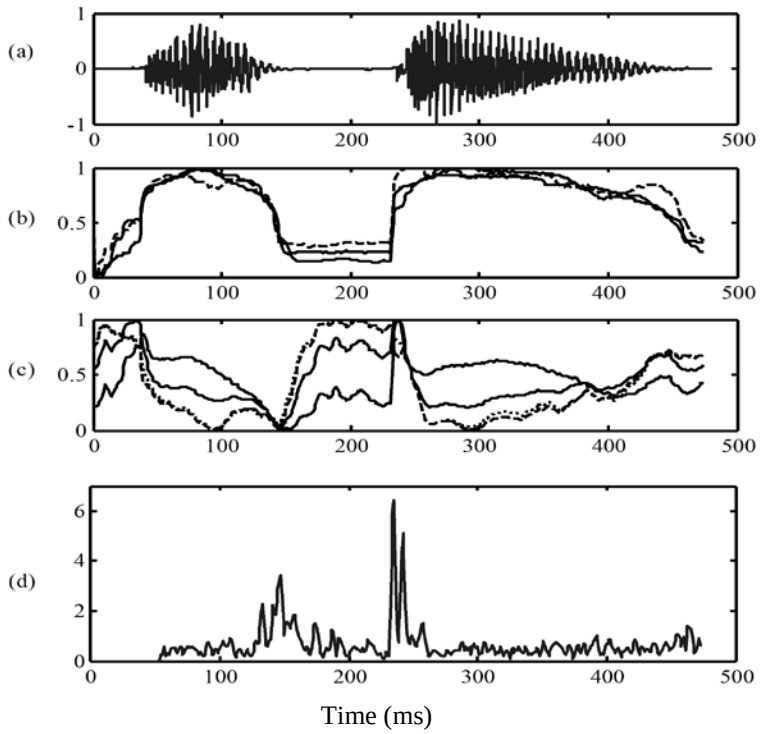


Figure 3.11 Landmark detection method SM for /aba/: (a) signal waveform, (b) band energy parameters E_{b1} (thick), E_{b2} (thin), E_{b3} (dashed) (c) spectral moments F_c (thick), F_σ (thin), F_s (dashed), F_k (dotted), (d) ROC-MD.

3.4.3 Evaluation of the SM method

The method was evaluated using the speech material consisting of VCV utterances and TIMIT sentences as used for the earlier methods and described in Section 3.2.2. Voicing offsets (−g) and onsets (+g) were located by the method described earlier in Section 3.2.1. A total of seven parameters, three band energies (E_{b1}, E_{b2}, E_{b3} from the bands 1.2 − 2.0, 2.0 − 3.5, 3.5 − 5.0 kHz, respectively) and four spectral moments (F_c, F_σ, F_s, F_k), were used for burst onset (b) detection. ROC-MD of band energies and spectral moments were computed individually and in a combined fashion for time-steps of 3 and 6 ms. The location of the most prominent peak in the ROC-MD track was taken as the burst onset landmark. Detection rates of voicing offset and onset landmarks are listed in Table 3.7. The burst onset detection rates at the temporal accuracy of 30, 20, 15, 10, 7, and 5 ms are listed in Table 3.8 for time-steps of 3 and 6 ms. The detection rates were generally lower for the larger time-step. Compared to spectral moments, band energy parameters contributed more towards burst onset detection. When used with energy parameters, spectral moments improved the burst onset detection rates.

In the evaluation using TIMIT sentences, location of the most prominent peak in the ROC-MD track within a [–g, +g] region was taken as the burst onset landmark. A valid closure interval was defined by a minimum duration of 10 ms with energy 20 dB below the

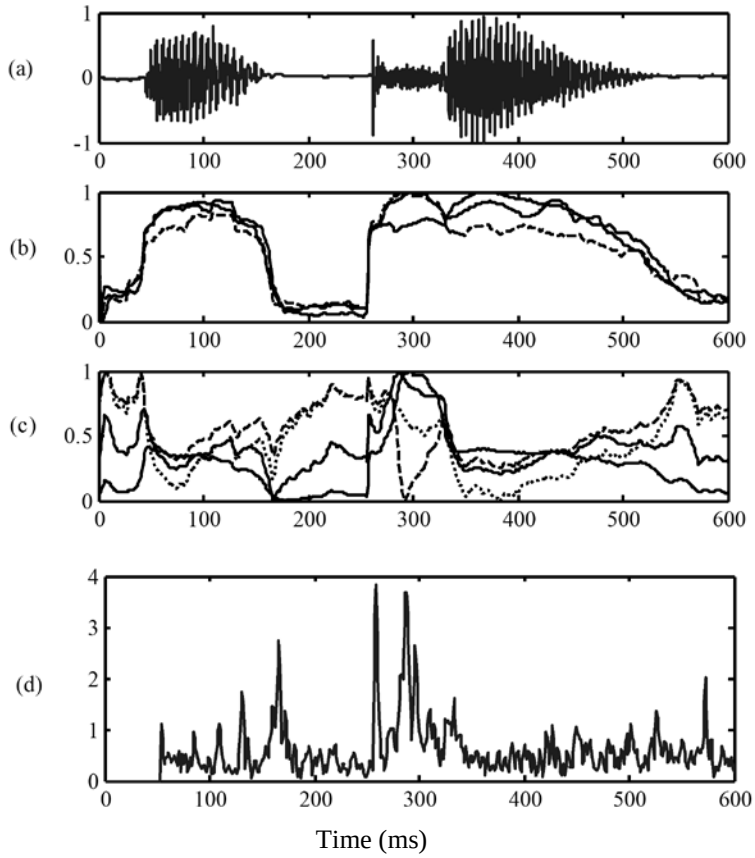


Figure 3.12 Landmark detection method SM for /apa/: (a) signal waveform, (b) band energy parameters E_{b1} (thick), E_{b2} (thin), E_{b3} (dashed), (c) spectral moments F_c (thick), F_σ (thin), F_s (dashed), F_k (dotted), (d) ROC-MD.

vowel energy in the utterance. In case of unpaired −g and +g, the search interval was limited to [−g, −g+50] and [+g−50, +g], respectively. The detection rates for burst onset, voicing offset, and voicing onset landmarks at temporal accuracy levels of 30, 20, 15, 10, 7, and 5 ms are listed in Table 3.9. The method also detected silence to vowel/semivowel onsets, frication onsets, glottal stops/clicks, with the insertion rates being 11, 9, and 4% respectively.

The use of spectral moments as parameters for burst onset detection and the use of Mahalanobis distance based rate of change was investigated. Energy parameters were found to be reliable and to significantly contribute towards detection. Spectral moments were useful as additional parameters for improving detection rates. Rate of change obtained by Mahalanobis distance based first difference (ROC-MD) was effective in combining the parameter variations and giving a single parameter indicative of the overall variation. The use of short time-steps of the order of 3 ms performed better for localizing burst onsets. The variations in the spectral centroid and higher order moments during frames with very low energy such as closure intervals was the main limitation of this approach.

Table 3.7 Landmark detection method SM: Detection rates (%) for voicing offsets (−g) and onsets (+g) in VCV utterances.

Landmarks (no. of tokens)	Temporal accuracy (ms)					
	≤30	≤20	≤15	≤10	≤7	≤5
−g (180)	93	83	73	64	56	36
+g (180)	99	98	98	93	83	76

Table 3.8 Landmark detection method SM: detection rates (%) for burst onset landmarks in VCV utterances.

Parameter set	Time step K	Temporal accuracy (ms)					
		≤30	≤20	≤15	≤10	≤7	≤5
$E_{b1,}E_{b2},E_{b3}$	3	97	97	96	93	93	91
	6	97	96	95	93	93	85
$F_{c,}F_{\sigma},F_{s},F_{k}$	3	90	90	87	83	81	81
	6	76	76	73	72	71	66
$E_{b1,}E_{b2},E_{b3},$ $F_{c,}F_{\sigma},F_{s},F_{k}$	3	99	98	98	96	95	95
	6	99	99	99	96	95	90

Table 3.9 Landmark detection method SM: Detection rates (%) for stop consonant landmarks in TIMIT sentences.

Landmarks (no. of tokens)	Temporal accuracy (ms)					
	≤30	≤20	≤15	≤10	≤7	≤5
−g (270)	90	80	63	40	27	19
+g (232)	96	91	82	71	60	45
b (306)	93	87	86	71	66	59

3.5 Method based on spectral moments with tone addition (SMTA)

During low energy segments such as stop closures, the spectral centroid has random variations and these variations affect the higher order moments. To stabilize the estimated values during low energy segments without significantly affecting them elsewhere, a low-level low-frequency tone was added to the speech signal. The effect of addition of tones at different frequencies and levels was investigated. It was observed that tones of frequency above 0.5 kHz suppressed the change in centroid during the closure-to-burst onset transition. Low frequency tones were effective in simulating the effect of a voice bar during the closure interval. However, use of tones of frequency below 100 Hz was not effective because of the analysis window duration of 6 ms. Evaluations were made with sinusoidal tone of frequency 100 Hz added at the levels of −40, −30, −20, −10, and 0 dB with respect to the maximum signal level. As examples of processing, Figure 3.13 and Figure 3.14 show the centroid tracks

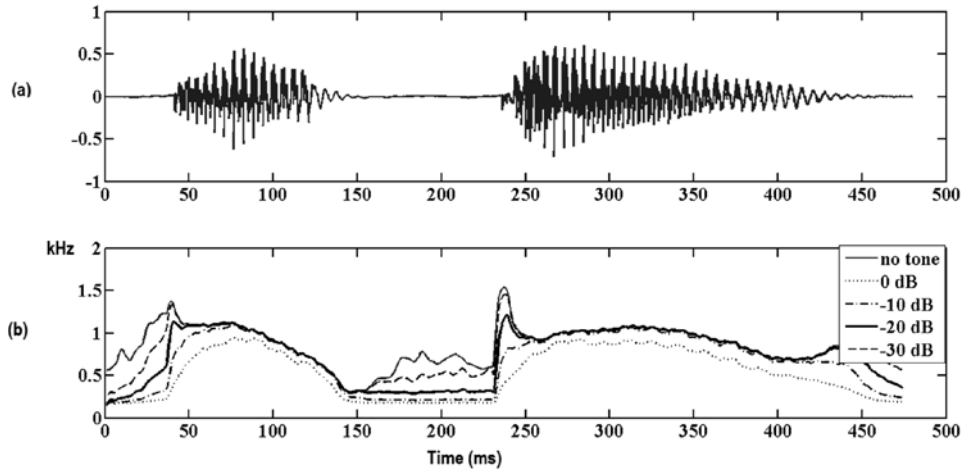


Figure 3.13 Landmark detection method SMTA for /aba/: (a) signal waveform, (b) centroid (thin), centroid computed from signal with tone added at: 0 dB (thin dotted), −10 dB (dash-dot), −20 dB (thick solid), −30 dB (thin dashed).

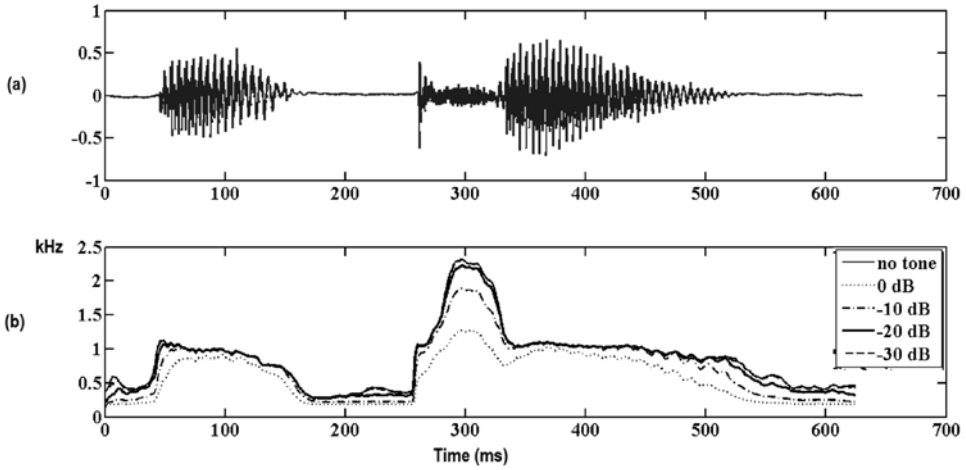


Figure 3.14 Landmark detection method SMTA for /apa/: (a) signal waveform, (b) centroid (thin), centroid computed from signal with tone added at: 0 dB (thin dotted), −10 dB (dash-dot), −20 dB (thick solid), −30 dB (thin dashed).

computed from signal with tone added at different levels. Addition of tones at −40 and −30 dB resulted in almost no change in the parameter tracks, while that at −10 dB and higher masked the abrupt transitions associated with the burst onsets. Addition of the tone at −20 dB resulted in stabilization of tracks during low energy segments without smearing the transitions associated with burst onsets. Hence in the revised landmark detection method based on spectral moments with tone addition (SMTA), a 100 Hz tone at −20 dB with respect to the maximum signal level was added to the signal before calculating the spectral moments.

The band energy parameters E_{b1}, E_{b2}, and E_{b3} from the bands 1.2 − 2.0, 2.0 − 3.5, and 3.5 − 5.0 kHz, were computed from the smoothed magnitude spectrum $X(n,k)$ using

(3.1). The smooth magnitude spectrum $X_t(n,k)$ of the tone-added signal was obtained using the method described in section 3.2.1. The first four spectral moments (F_{ct}, $F_{\sigma t}$, F_{st}, F_{kt}) were computed from the smoothed magnitude spectrum $X_t(n,k)$ using the method described earlier in section 3.4.1. Mahalanobis distance based ROC was computed for the band energy parameters E_{b1}, E_{b2}, E_{b3} and spectral moments $F_{ct}, F_{\sigma t}, F_{st}, F_{kt}$ individually and in a combined fashion. The location of the most prominent peak in the ROC-MD track was taken as the burst onset landmark.

Evaluation was performed using 180 VCV utterances involving 6 stop consonants (/b, d, g, p, t, k/) in three vowel contexts (/a, i, u/) recorded from 10 speakers (5 male and 5 female) as used in the earlier methods. The results are listed in Table 3.10. It is seen that the centroid F_{ct} contributed most, followed by the band energy parameter E_{b1}. Out of the different parameter sets, the highest detection rates were obtained using (E_{b1}, F_{ct}). It is further noted that inclusion of higher moments ($F_{\sigma t}, F_{st}, F_{kt}$) decreased the detection rate. Hence further investigation is carried out using only centroid as the parameter. The effect of tone addition was investigated further using 50 sentences from TIMIT database as used for the earlier methods and described in Section 3.2.2. The centroids F_c and F_{ct} with and without tone addition were computed for all the sentences. The rate of change (ROC) functions of F_c and F_{ct} were computed using a time-step *K* of 20 ms as

$$dF_c(n) = F_c(n) - F_c(n-K) \tag{3.26}$$

$$dF_{ct}(n) = F_{ct}(n) - F_{ct}(n-K) \tag{3.27}$$

From the TIMIT transcription files corresponding to the 50 sentences, the phonemes were grouped as stops, fricatives, nasals, semivowels, and vowels. The values of dF_c and dF_{ct} were obtained from the corresponding contours at the locations in the TIMIT transcription files for all these phone classes. The means and standard deviations of dF_c and dF_{ct} are given in Table 3.11. The onsets of stops and fricatives are associated with large positive values of the mean dF_c. The onsets of vowels, semivowels, and nasals are associated with comparatively small positive or negative values. Mean values of dF_{ct} show that the addition of tone increases the mean value for onsets of stop release bursts almost by a factor of 2 and only slightly decreases the value for fricatives. Standard deviations are not much affected by tone addition. Thus dF_{ct} is likely to be more effective than dF_c in the detection of burst and frication onsets. Burst and frication onset landmarks are detected by comparing dF_{ct} with an empirically set threshold.

The method was evaluated using 50 conversational style sentences (3 female and 2 male speakers × 10 sentences) from TIMIT database. The detection rates at temporal accuracies of 30, 20, 15, 10, 7, and 5 ms for stop release bursts and frication onsets are listed in Table 3.12. A threshold value of 350 Hz was used in the investigation. The method

Table 3.10 Landmark detection method SMTA: Detection rates (%) for burst onset landmarks in VCV utterances using different sets of parameters.

Parameter set	Temporal accuracy (ms)					
	≤20	≤15	≤10	≤7	≤5	≤3
E_{b1}	95.3	94.7	94.7	93.6	92.4	91.8
E_{b1}, E_{b2}	96.5	94.7	94.7	93	91.2	90.1
E_{b1}, E_{b2}, E_{b3}	95.9	94.2	92.4	90.6	88.9	87.7
F_{ct}	100	100	98.2	94.7	93.6	88.3
F_{ct}, $F_{\sigma t}$	98.2	97.1	94.7	93.6	91.8	90.1
F_{ct}, $F_{\sigma t}$, F_{st}	78.4	76.6	71.9	68.4	65.5	62.0
F_{ct}, $F_{\sigma t}$, F_{st}, F_{kt}	83.0	81.3	76.0	72.5	70.2	65.5
E_{b1}, F_{ct}	99.4	98.8	98.8	97.1	95.9	94.7

Table 3.11 Landmark detection method SMTA: Mean and standard deviation (std.) of $d_{Fc}(n)$ and $d_{Fct}(n)$ at the onsets for different phoneme classes in TIMIT sentences.

Phoneme class (no. of tokens)	Without tone addition, dF_c		With tone addition, dF_{ct}	
	Mean	Std.	Mean	Std.
Stops (306)	505	533	1119	689
Fricatives (194)	1220	788	889	658
Nasals (162)	−87	224	−72	202
Semivowels (211)	−33	300	57	331
Vowels (623)	35	420	203	484

Table 3.12 Landmark detection method SMTA: Detection rates (%) for onsets of burst and frication in TIMIT sentences.

Landmarks (no. of tokens)	Temporal accuracy (ms)					
	≤30	≤20	≤15	≤10	≤7	≤5
Burst (306)	88	86	75	65	51	24
Frication (194)	89	86	75	60	39	24

detected 86% of stop release bursts and frication onsets with a temporal accuracy of 20 ms. The method had an insertion rate of nearly 19% which involved semivowel to vowel (8%) and nasal to vowel (11%) transitions.

An evaluation of burst onset detection using the method SMTA was performed using 180 VCV utterances and 300 CVC keywords in the MRT utterances. The stop release bursts associated with the six stop consonants (/b, d, g, p, t, k/) in the VCV utterances and CVC keywords were manually marked by observing the waveforms and spectrograms. There were

Table 3.13 Landmark detection method SMTA: Detection rates (%) at different temporal accuracies (ms) for stop release bursts in nonsense VCV syllables with stop consonants /b, d, g, p, t, k/ and vowels /a, i, u/.

Stop	No. of tokens	Temporal accuracy (ms)					
		≤30	≤20	≤15	≤10	≤7	≤5
b	30	97	97	97	93	90	87
d	30	100	100	100	100	100	100
g	30	97	97	97	97	93	93
p	30	100	100	100	100	93	93
t	30	100	100	100	97	90	90
k	30	100	100	100	100	97	97
Overall	180	99	99	99	98	92	92

Table 3.14 Landmark detection method SMTA: Detection rates (%) at different temporal accuracies (ms) for stop release bursts in MRT utterances.

Stop	No. of tokens	Temporal accuracy (ms)					
		≤30	≤20	≤15	≤10	≤7	≤5
b	35	69	60	54	40	31	26
d	42	95	95	95	93	81	69
g	31	97	97	94	94	94	90
p	56	69	69	68	68	65	64
t	78	99	99	99	97	95	90
k	63	89	87	86	79	68	57
Overall	305	87	86	84	81	74	68

180 stop release bursts in the VCV utterances and 305 stop release bursts in the CVC words. The crossing point of $dF_c(n)$ exceeding a threshold frequency of 300 Hz was taken as the location of the burst onset. The detection rates of release bursts for six stop consonants at different temporal accuracies for VCV utterances and CVC words are listed in Table 3.13 and Table 3.14, respectively. For VCV utterances, 99% of the stop release bursts were detected within 20 ms of the manually marked burst locations, and 92% were detected within 5 ms of the manually marked locations. For CVC words, the overall detection rates were 86, 81, and 68% at temporal accuracy of 20, 10, and 5 ms, respectively. The lower detection rates for CVC words may be due to a more frequent occurrence of weak and unreleased stop release bursts, particularly in the word final position.

The detection rates for labial stop release bursts were lower than that for alveolar and velar bursts. This could be due to the comparatively low energy and centroid frequencies of labial stops compared to velar and alveolar stops. In addition to stop release bursts, the method detected 71% of onsets of voiced and unvoiced fricatives (a total of 80 tokens) with a

temporal accuracy of 20 ms. The method occasionally detected transitions associated with glides in CVC words, but it may not adversely affect the enhancement in speech intelligibility to be gained by CVR modification. The detection rates for the method were comparable to the energy and centroid based method (Jayan and Pandey, 2012), and about 4% higher than the detection rates of the method used by Colotte and Laprie (2000) for locating regions for modification for speech intelligibility enhancement.

3.6 Discussion

From the evaluations of landmark detection schemes, it can be concluded that the performance of the landmark detector depends on the parameter set used for capturing the spectral variations and the method used for quantifying the parameter variations. Band energy parameters from spectral bands with fixed boundaries have low adaptability to the variability of speech. Their usability in precisely detecting the acoustically abrupt events such as burst onsets seems to be limited due to the asynchronous nature and the variable dynamic range of band energy variations during the abrupt spectral transitions. Parameters which capture the overall spectral shape variations improve the temporal accuracy of landmark detection. Gaussian parameters modeling the short-time speech spectrum were found useful for detecting the burst onset landmarks. However, the method is unsuitable for real-time implementation because of the computation intensive parameter estimation and the use of large algorithmic delays for landmark detection. Mahalanobis distance based rate of change measure was found to be effective in combining variations in parameters with different magnitude scales and correlations and to give a single parameter indicative of the overall variation. The use of short time-steps of the order of 3 ms helped in precise time-localization of the burst onset landmarks. The addition of sinusoidal tone to the speech signal at a level of −20 dB with respect to the maximum vowel level stabilized the random variations in centroid during the closure intervals and improved the detection rates of burst onset landmarks.

In the three methods (EC, GMM, SM) investigated for detection of stop consonant landmarks, burst onset detection was based on localizing the most abrupt change towards higher frequencies in the regions bounded by voicing offset and onset separated by a closure interval. This approach involves an algorithmic delay of the order of 400 ms, which is infeasible in a real-time implementation. These methods have comparatively higher detection rates and are more suited for speech intelligibility enhancement applications involving offline processing. To reduce the algorithmic delays, we need to detect the landmarks of interest in a single pass with limited contextual information, preferably using a robust parameter variation. The spectral centroid computed from the signal added with tone seems to be a good candidate for real-time burst onset detection. Centroid parameter is an indicative of the spectral shape and is relatively independent of the signal level. For a sampling frequency of f_s, centroid is

always in the range of 0 to 0.5 f_s. This gives an advantage for centroid over the band energy parameter for real-time detection of burst onsets as the burst and non-burst transitions can be well separated by the use of a threshold. The method SMTA has relatively low computational complexity and a small algorithmic delay, as the detection involves comparison of rate of change of centroid with an empirically set threshold. The insertions, mainly caused by the detection of frication onsets may not pose a problem for intelligibility enhancement using CVR modification as the method is reported to be equally effective for stops and fricatives.

Chapter 4

AUTOMATED ENHANCEMENT OF SPEECH INTELLIGIBILITY

4.1 Introduction

Automated transformation of conversational speech to clear speech by signal processing is very difficult due to the segment-specific and non-uniform nature of the acoustic differences between them. The acoustic landmarks with concentration of perceptual cues are pronounced in clear speech, but may be reduced or missing in conversational speech due to the increased co-articulation effects. This makes the automated detection and modification of landmarks difficult and challenging in conversational speech. The perceptual effects of careful and increased articulation efforts in clear speech cannot be fully recreated by signal processing of conversational speech. For example, the well defined formant targets achieved during clear speech production cannot be achieved by mere time-scale modification of conversational speech. The use of uniform scaling factors for speech characteristic modification and the perceptual distortions introduced during modification limit the effectiveness of automated signal processing schemes. Further, the excessive computational requirements and processing delays involved in many possible signal processing solutions restrict their usefulness for real-time enhancement of speech intelligibility in speech communication devices and hearing aids.

This chapter presents investigations on enhancement of speech intelligibility by automated CVR and time-scale modification. An offline algorithm described earlier in Section 3.4 is used for the detection of regions for modification. Because of the improved detection rates of the offline algorithm, this approach is expected to bring out the possible benefits that could be derived from CVR and time-scale modification. This method is suited for processing of pre-recorded speech material for improving perception by hearing-impaired listeners or by normal-hearing listeners in adverse listening conditions. It can also be used as a pre-processing stage for improving performance of systems involving automated speech recognition.

Two experiments were conducted using isolated nonsense VCV utterances involving 6 stop consonants (/b, d, g, p, t, k/) paired with 3 vowels (/a, i, u/) recorded from 5 speakers as the test material. Experiment I involved CVR modification and Experiment II involved time-scale modification. The use of isolated nonsense utterances as test material helped to study the

effects of signal processing on perception of stop consonants with minimal semantic and co-articulation effects.

4.2 Signal processing for automated CVR modification

CVR modification involved amplification of VC transition, CV transition, and the stop release burst. In this processing, the voicing offset (−g), burst onset (b), and voicing onset (+g) landmarks are located using an automated landmark detection method based on spectral moments as described earlier in Section 3.4. The regions selected for CVR modification correspond to [−g−20, −g] for VC transition and to [b−10, +g+10] for CV transition, respectively, with −g, b, +g being detected landmark locations in ms. This results in amplification of approximately 2 pitch cycles (20 ms) during the VC transition. In the CV transition, the amplification starts 10 ms before the burst onset and continues for approximately 1 pitch cycle after the voicing onset. Amplitude discontinuity during CVR modification is avoided using a cosine tapered window with rise and fall times of 2.5 ms. Informal listening tests conducted using VCV utterances with six stop consonants (/b, d, g, p, t, k/) in the three vowel contexts (/a, i, u/) as test material showed CVR modification by 9 dB to be most effective in terms of improving the audibility of the consonants without introducing perceptible distortions. As examples of processing, the waveforms, boundaries selected and scaling functions used for CVR modification, and the CVR modified waveforms for VCV utterances /aga/ and /aka/ are shown in Figures 4.1 and 4.2, respectively.

4.3 Signal processing for automated time-scale modification

The effect of expansion of transition segments on the identification of stop consonants was investigated earlier using an automated method (Jayan et al., 2007; 2008). The segment boundaries for time-scale expansion were located using a transition index derived from spectral band energies and centroids. The signal processing for time-scale modification was performed using harmonic plus noise model (HNM) (Laroche et al.,1993; Stylianou, 2001; 2005; Pantazis and Stylianou, 2008). VCV utterances involving 6 stop consonants (/b, d, g, p, t, k/) paired with vowel /a/ were used as the test material. The VC and CV transitions were expanded by a time-scaling factor and the steady-state vowel segments on either side were appropriately time compressed, maintaining the overall speech duration unaltered. Modifications were carried out with time-scaling factors of 1.0, 1.2, 1.5, 1.8, and 2.0. Evaluation in the presence of broadband noise at different SNRs (0, −3, −6, −9, −12 dB) indicated time-scaling factor of 1.5 to be optimum for the expansion of transition segments. The signal processing for time-scale modification occasionally introduced audible distortions, possibly due to improper fusion of the harmonic part with the noise part during synthesis.

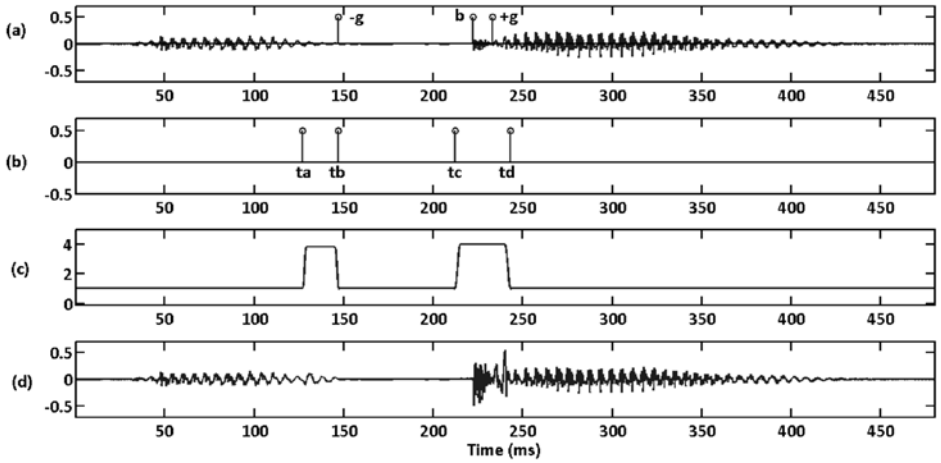


Figure 4.1 CVR modification of VCV utterance /aga/: (a) waveform with landmarks, (b) boundaries of windows selected for CVR modification, (c) scaling function for CVR modification, (d) modified waveform.

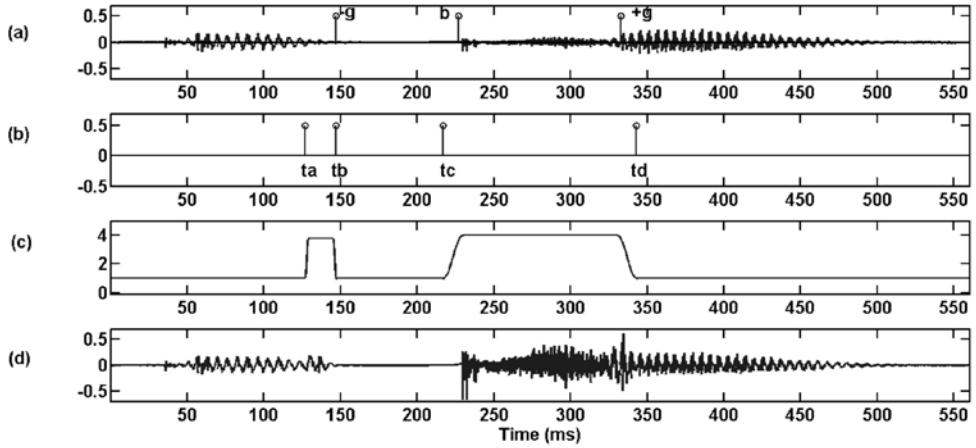


Figure 4.2 CVR modification of VCV utterance /aka/: (a) waveform with landmarks, (b) boundaries of windows selected for CVR modification, (c) scaling function for CVR modification, (d) modified waveform.

During transition and low energy segments, the errors in the estimation of pitch and model parameters may adversely affect the quality of HNM based synthesis (Pantazis and Stylianou, 2008). To reduce the processing related artifacts during time-scale modification, the use sinusoidal model based analysis-modification-resynthesis (McAulay and Quatieri, 1986) was investigated. Sinusoidal model based processing has been used for spectrum-invariant time-scale and pitch-scale modifications of speech signal (Quatieri and McAulay, 1992). Kates (1994) investigated sinusoidal model based spectral contrast enhancement for improving speech intelligibility and reported consonant recognition to be related to the number of sinusoids used in the modeling process. Vowels and consonants with compact spectra could

be modeled with parameters related to 16 prominent spectral peaks, whereas consonants with diffused spectra needed more number of sinusoids for proper representation.

4.3.1 Sinusoidal model based analysis-synthesis

In sinusoidal model based analysis-synthesis (McAulay and Quatieri, 1986), the speech signal is modeled as a sum of sinusoidal components of time-varying amplitudes, frequencies, and phases. Frequencies of sinusoids are harmonically related for voiced speech, but are generally unrelated for unvoiced speech. Let speech signal $s(i)$ in the *n*th frame be modeled by L_n sinusoidal components with amplitudes A_{nk}, frequencies ω_{nk}, and phase offsets ϕ_{nk} and be given as the following

$$s(i) = \sum_{k=1}^{L_n} A_{nk} \cos(\theta_{nk}(i)) \qquad (-N/2 \leq i < N/2) \tag{4.1}$$

where *N* is the frame length, *k* and *i* are the indices for the sinusoidal component and sample number, respectively. The phase of the *k*th component is given in terms of frequency and phase offset as

$$\theta_{nk}(i) = \omega_{nk} i + \phi_{nk} \tag{4.2}$$

A block diagram representation of the sinusoidal model based analysis is shown in Figure 4.3. A close approximation of the frame $s(i)$ is obtained by picking L_n prominent peaks in the magnitude spectrum. The angular frequencies corresponding to the peak locations are taken as ω_{nk} and the spectral amplitudes and phases at the peak locations are taken as A_{nk} and ϕ_{nk}, respectively (McAulay and Quatieri, 1986). To get adequate frequency resolution during the peak picking process, the analysis window length is kept higher than two pitch periods.

A block diagram representation of the sinusoidal synthesis is shown in Figure 4.4. During synthesis, the parameters estimated for individual frames are interpolated across frames to eliminate discontinuities at the frame boundaries. Sinusoids of consecutive frames are matched based on their frequencies to form frequency tracks using a nearest neighbor algorithm. For each frequency track, amplitude and phase tracks are obtained by linear and cubic interpolation, respectively (McAulay and Quatieri, 1986). The synthesized signal $s'(i)$ for *n*th frame is obtained as

$$s'(i) = \sum_{k=1}^{L_n} B_{nk}(i) \cos(\varphi_{nk}(i)) \tag{4.3}$$

where $B_{nk}(i)$ and $\varphi_{nk}(i)$ are the values of the amplitude and phase tracks of the *k*th sinusoid at sample *i*.

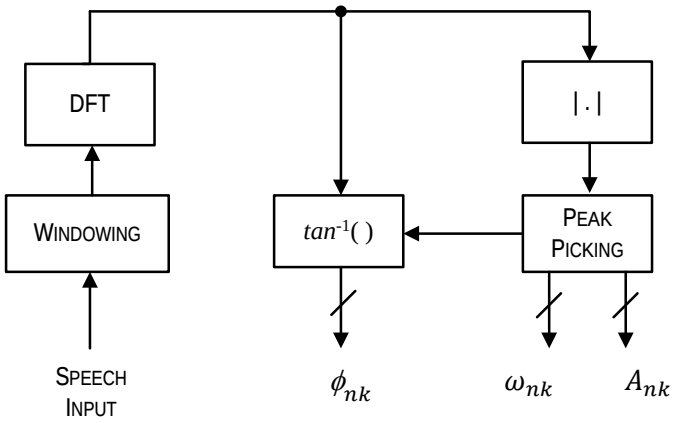


Figure 4.3 A block diagram representation of sinusoidal model based analysis (McAulay and Quatieri, 1986).

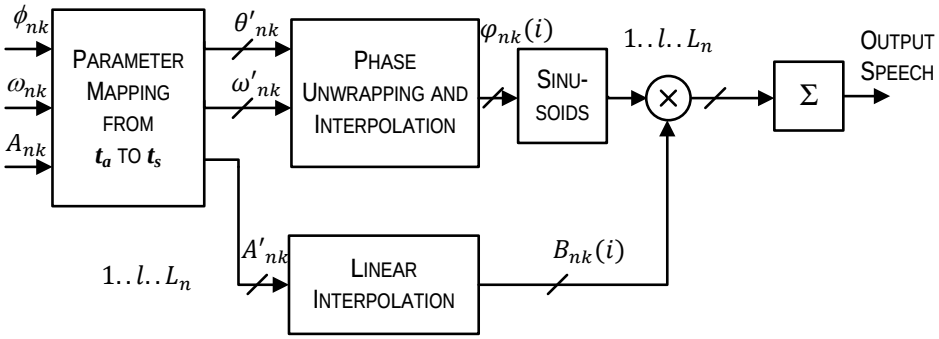


Figure 4.4 A block diagram representation of sinusoidal model based synthesis (McAulay and Quatieri, 1986).

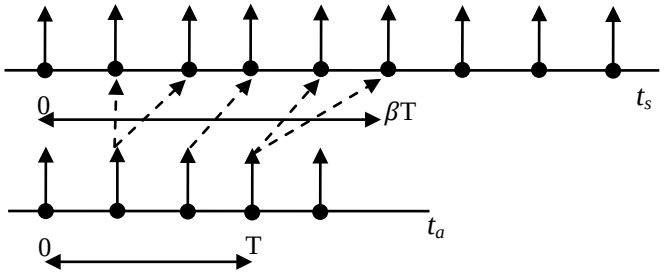


Figure 4.5 Time-scale modification: mapping on onset points for β = 1.5.

4.3.2 Time-scale modification

The objective of time-scale modification is to alter the rate of articulation of speech by stretching or compressing the time-evolution of the formant structure and the pitch contour (Quatieri and McAulay, 1986; Moulines and Laroche, 1995). The time instant t_a in the analysis time axis gets mapped to a time instant t_s in the synthesis time axis, where

$$t_s = \beta t_a \tag{4.4}$$

β being the scaling factor. The rate of articulation is reduced during time-scale expansion ($\beta > 1$), and increased during time-scale compression ($\beta < 1$). The disruption of phase relationships across sinusoids of different frequencies at the synthesis time instants may cause perceptible distortion during time-scale modification. This problem is reduced by performing the analysis and synthesis in a pitch-synchronous manner with frames centered at absolute or relative onset times (Quatieri and McAulay, 1986).

The parameters along the analysis time instants t_a are mapped to the synthesis time instants t_s in accordance with the time-scaling factor β = 1.5 as given in Figure 4.5. During time-scale expansion, the parameter set of alternate analysis time instants are duplicated at the newly introduced synthesis time instants. Amplitude and unwrapped phase parameter tracks are obtained along the synthesis time axis by linear and cubic interpolation, respectively. The time-scale modified speech is obtained using (4.3) using the new parameter tracks.

A sinusoidal model based analysis-modification-resynthesis platform was developed in Matlab. For a sampling frequency of 10 kHz, a model with 80 sinusoids estimated from 512-point DFT on 20 ms frames has been reported to be sufficient for reproduction of voiced and unvoiced speech with quality almost indistinguishable from that of the original (McAulay and Quatieri, 1986). Our implementation is based on 160 sinusoids estimated using 1024-point DFT on 20 ms Hamming windowed frames. During analysis, the frames were placed pitch synchronously during voiced segments and with a fixed separation of 5 ms during unvoiced segments. An estimate of pitch and locations of frames during voiced segments were obtained using an algorithm proposed by Boersma and Weenink (1992). The burst onset (b) and the voicing onset landmarks (+g) were detected in an automated manner using the method based on spectral moments and described in Section 3.4. The segment starting at the burst onset and extending to approximately 5 pitch cycles after the voicing onset were expanded by time-scaling factor 1.5. Thus the region selected for time-scale modification extended from t_1 to t_2 where

$$t_1 = b \tag{4.5}$$

$$t_2 = +g + 50 \tag{4.6}$$

As examples of processing, time-scale modification of VCV utterances /aga/ and /aka/ are shown in Figures 4.6 and 4.7, respectively. Figures show the waveforms of unprocessed utterances, time-scaling factors used for expansion of the CV transition regions, waveforms of synthesized and the time-scale modified utterances using sinusoidal model based analysis/modification/synthesis. It is noticed that the synthesized utterance is of the same

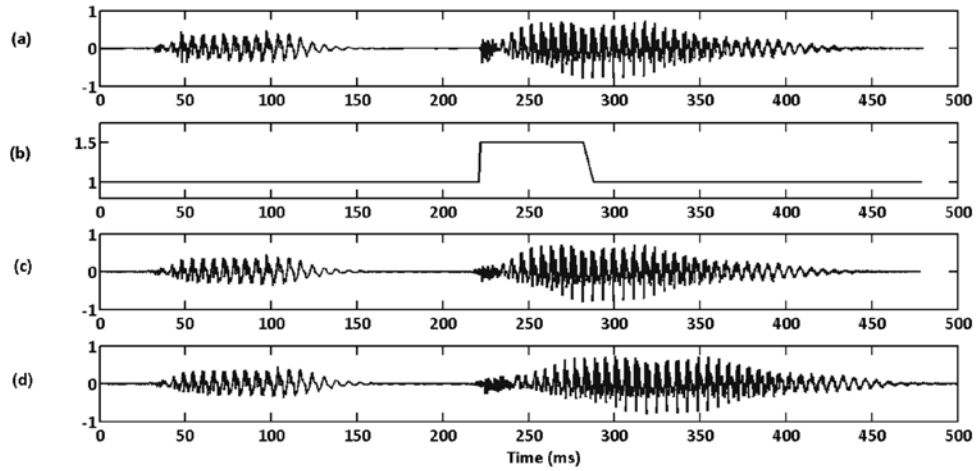


Figure 4.6 Time-scale modification of CV transition of /aga/: (a) unprocessed waveform, (b) time-scaling factor β for expansion of CV transition, (c) re-synthesized waveform, and (d) time-scale modified waveform.

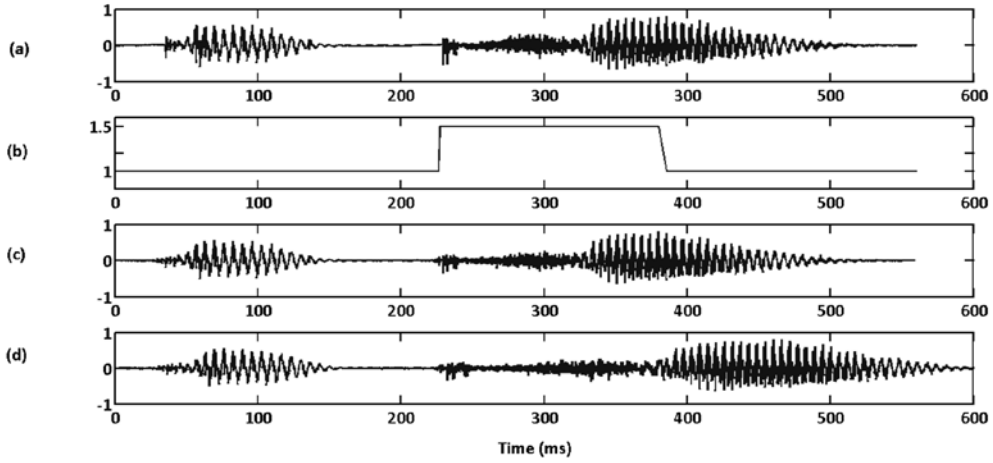


Figure 4.7 Time-scale modification of CV transition of /aka/: (a) unprocessed waveform, (b) time-scaling factor β for expansion of CV transition, (c) re-synthesized waveform, and (d) time-scale modified waveform.

duration as that of the original utterance and the time-scale modified VCV utterance is of a slightly longer duration with expanded CV transition.

4.4 Listening Tests

Two sets of listening tests involving closed-set recognition of VCV utterances were conducted using normal hearing subjects: Experiment I to evaluate the effect of CVR modification and Experiment II to evaluate the effect of time-scale modification. In both the experiments, speech-spectrum shaped noise (Hazan and Simpson, 1998; Liu, 1996; Yoo et al., 2007) with long-term spectrum nearly flat from 100 Hz to 1 kHz and 12 dB/octave roll-off

afterwards was used as the masking background. The SNRs ranged from −12 to 12 dB in steps of 6 dB.

4.4.1 Material

The test material consisted of 18 nonsense VCV utterances with 6 stop consonants (/b, d, g, p, t, k/) paired with 3 vowels (/a, i, u/) each recorded from 5 speakers (2 male and 3 female) at a sampling frequency of 10 kHz. Thus there were a total of 90 stimuli (18 utterances × 5 speakers). The stimuli were normalized to have the same RMS value during the vowel segments. In Experiment I, the stimuli were processed for 9 dB CVR modification. It involved amplification of the transition and burst segments and did not affect the vowel level in the utterances. Thus there were two types of stimuli: unprocessed (unp) and CVR modified (cvr). In Experiment II, there were three types of stimuli: unprocessed (unp), synthesized (syn), and time-scale modified by a factor 1.5 (tsc). Speech-spectrum shaped noise was subsequently added as a masker with respect to the RMS value of the vowel segment to get SNRs of ∞ (quiet), +12, +6, 0, –6, and –12 dB. The noise extended for 1 s on either side of the stimuli. Examples of the speech files used as the test material are available on the web (Jayan, 2014).

4.4.2 Method

Consonant recognition tests were conducted on five normal-hearing subjects (3 male and 2 female, 18−40 years) using a computerized test administration setup. In Experiment I, there were 12 listening conditions: 2 types of stimuli (unp: unprocessed, cvr: processed) × 6 SNRs. In Experiment II, there were 18 listening conditions: 3 types of stimuli (unp: unprocessed, syn: synthesized, tsc: time-scaled) × 6 SNRs. The audio amplifier gain was set by the listener for the most comfortable listening level for the loudest set of sounds, i.e. those corresponding to SNR of −12 dB. The same gain setting was used across the listening conditions to maintain the same level for the speech signal. The stimuli were presented through Sennheiser PX80 headphones. After each presentation, the listener responded by clicking on one of the six response choices corresponding to the six consonants displayed on the computer screen. The position of the response choices displayed on the computer screen was randomized to avoid position bias in the responses.

In a multiple choice test, the response time provides a measure of the load on the perception process (Baer et al., 1993; Apoux et al., 2001). It is an indicator of the effort involved in integrating the perceptual cues in recognizing the presented stimulus. An increase in the response time indicates an adverse listening condition. It is desirable that the processing for intelligibility enhancement should decrease the perceptual load or at least should not

increase it. Therefore in addition to recording the subject's response to the presented stimulus, the response time was also recorded.

In both experiments, the test for each listening condition for each subject involved 4 presentations of each of the 90 stimuli. The stimuli were presented in a randomized order and were distributed over six sessions, each session involving 60 presentations and taking about 10 minutes for completion. The order of listening conditions was randomized across the listeners to reduce the effects of practice and fatigue. The instructions given to the subjects and the forms for collection of background information of the subjects and their consent to participate in the tests are given in Appendix A.

The responses to the stimuli were tabulated as stimulus-response confusion matrices. These were used to obtain recognition scores as a function of SNR. The differences between the recognition scores for the unprocessed speech and those for the processed speech are indicators of enhancement in intelligibility due to processing at different SNRs. Increase in recognition scores because of processing may also be interpreted as an equivalent SNR advantage using a minimum mean square error calculation (Kapoor and Allen, 2012). For this purpose, the recognition scores $P_c(\mathrm{SNR}_k)$ at SNR values of −12, −6, 0, 6, and 12 dB were converted to error $P_e(\mathrm{SNR}_k) = 1 - P_c(\mathrm{SNR}_k)$ and the MATLAB procedure 'lsqcurvefit ()' was used to fit the following sigmoid function on it,

$$P_e(\mathrm{SNR}_k) = e_c/\left[1 + \exp\left(\lambda(\mathrm{SNR}_k - \mathrm{SNR}_0)\right)\right] \tag{4.7}$$

where e_c = 5/6 is the chance error (corresponding to 6 response options for each presentation). The parameter λ is a scaling factor and it is returned by the sigmoid curve fitting function 'lsqcurvefit' of MATLAB. It has a range of 0 – 1 and its value is dependent on the input data for curve fitting. The parameter SNR_0 is the SNR at which the recognition score becomes 50%. With sigmoid curves fitted to the data, SNR shift to be given to the curve for the processed sounds to get the best possible approximation to the curve for the unprocessed sounds is taken as the equivalent SNR advantage. The stimulus-response confusion matrices were also used to carry out information transmission analysis (Miller and Nicely, 1955) to get the relative information transmitted in terms of overall, voicing, and place features.

4.5. Results for CVR modification (Exp. I)

The recognition scores for the individual subjects along with the means and standard deviations are given in Table 4.1 and a plot of the mean scores is shown in Figure 4.8. The recognition scores for the unprocessed and the CVR modified stimuli were almost the same as the SNR decreased from ∞ to 6 dB. For SNRs of 0, –6, and –12 dB, the recognition scores of the unprocessed stimuli were 81, 68, and 41%, respectively. The corresponding values for the

Table 4.1 CVR Modification (Exp. I): Recognition scores (%) for VCV utterances (unp: unprocessed, cvr: CVR modified). *p*: significance level of one-tailed paired t-test.

Subj.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	97	96	96	94	93	94	79	94	57	78	35	56
SC	97	95	89	89	87	86	81	86	75	89	50	79
SK	86	87	84	84	83	85	80	86	67	84	37	63
SB	97	98	97	98	97	97	91	96	72	95	46	68
SN	87	87	84	84	80	80	76	82	67	84	37	63
Avg.	93	93	90	90	88	88	81	88	68	86	41	66
s.d.	5.8	5.2	6.3	6.2	7	6.9	5.7	5.9	6.8	6.4	6.6	8.5
Impr.		0		0		0		7		18		25
p								0.01		<0.001		<0.001

Table 4.2 CVR Modification (Exp. I): Recognition scores (%), averaged across subjects, for different vowel contexts.

Vow. cont.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
a	92	92	91	90	88	89	82	87	68	87	41	69
i	94	93	90	90	88	88	81	89	66	86	43	65
u	93	93	89	89	89	88	81	90	68	84	40	64

Table 4.3 CVR Modification (Exp. I): Recognition scores (%), averaged across subjects, for different stops.

Stop	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
b	86	85	77	72	75	73	70	71	64	72	42	57
d	94	96	94	94	90	94	90	97	80	97	46	79
g	92	85	89	89	87	87	74	87	57	87	25	46
p	92	93	93	92	91	89	81	90	61	79	45	64
t	98	100	97	99	98	99	94	99	77	99	55	92
k	96	96	89	89	88	89	78	88	66	81	34	56

CVR modified stimuli were 7, 18, and 25% higher. The improvements were statistically significant ($p < 0.001$) at −6 and −12 dB SNRs. Hazan and Simpson (1998) reported improvements of 6 and 12% in recognition scores for VCV utterances in speech-spectrum shaped noise background at SNRs of 0 and −5 dB, respectively. Compared with their results, our CVR modification method resulted in nearly 6% higher improvement at the lower SNR.

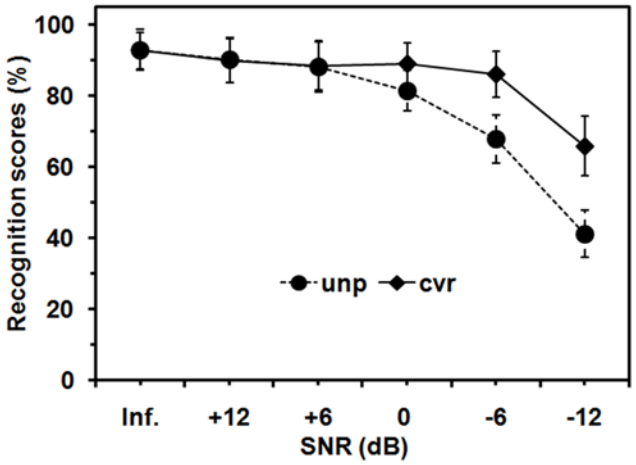


Figure 4.8 CVR Modification (Exp. I): Recognition scores (%) for VCV utterances *vs* SNR. Error bars indicate standard deviations.

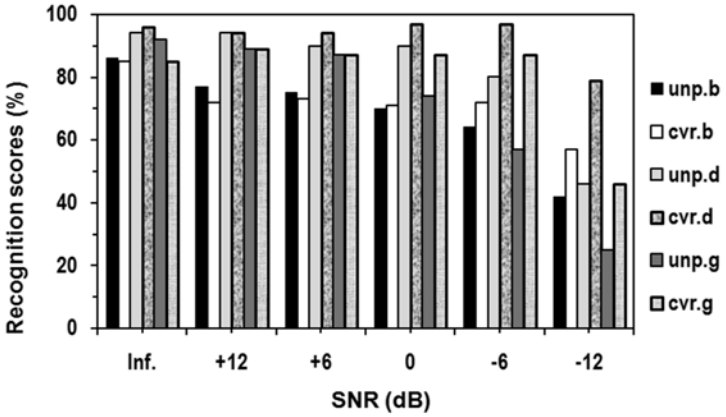


Figure 4.9 CVR Modification (Exp. I): Recognition scores (%) for voiced stops.

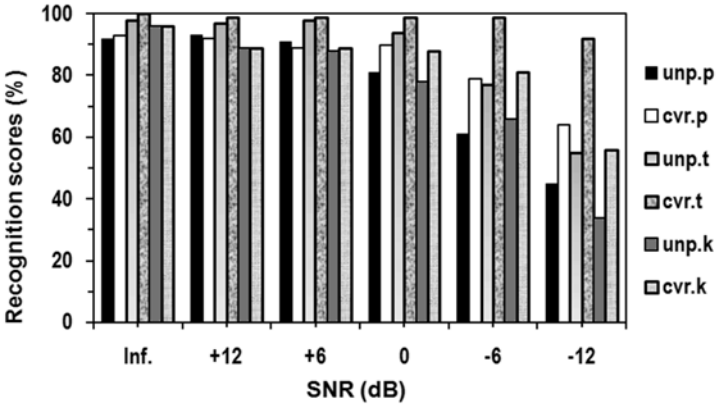


Figure 4.10 CVR Modification (Exp. I): Recognition scores (%) for unvoiced stops.

The increase in recognition scores averaged across the subjects corresponded to an SNR advantage of 6 dB (measured using the method as described earlier).

Table 4.4 CVR Modification (Exp. I): Relative information transmission (%). unp.: unprocessed stimuli, cvr.: stimuli processed with CVR modification.

SNR (dB)	Overall		Place		Voicing	
	unp	cvr	unp	cvr	unp	cvr
∞	93	91	90	88	98	95
12	79	79	84	84	67	67
6	75	76	77	80	66	66
0	63	78	55	81	69	66
−6	45	72	29	70	61	69
−12	20	43	6	32	34	51

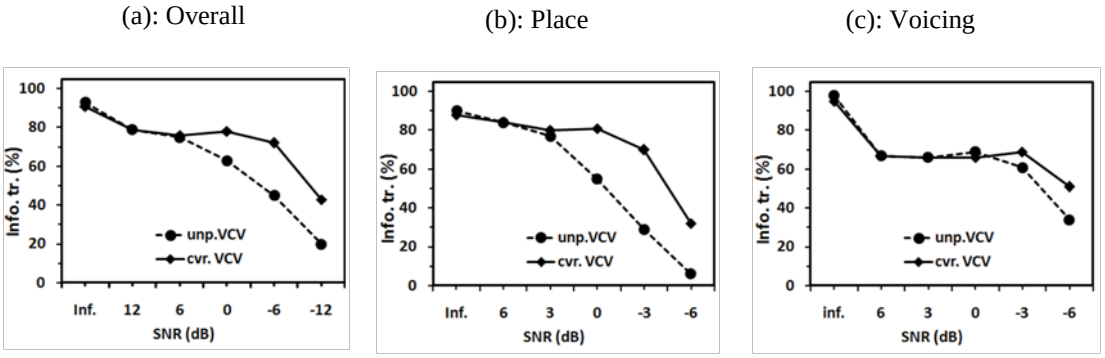


Figure 4.11 CVR Modification (Exp. I): Relative information transmission (%).

Recognition scores averaged across subjects for the three vowel contexts are listed in Table 4.2. The scores show that CVR modification is equally effective in the three vowel contexts. Consonant-wise recognition scores are given in Table 4.3 and these are plotted for voiced and unvoiced stop consonants in Figures 4.9 and 4.10, respectively. Compared with the labial (/b, p/) and velar stops (/g, k/), alveolar stops (/d, t/) were more benefitted by CVR modification at lower SNRs. Information transmission analysis (Miller and Nicely, 1955) was carried out on the stimulus-response confusion matrices of the individual subjects to get the relative information transmitted in terms of overall, voicing, and place features. The values averaged across the subjects are listed in Table 4.4 and plotted in Figure 4.11. It is seen that the loss in information at the lower SNRs was mainly due to the place feature and CVR modification was effective in improving its transmission.

Response time is an indicator of the perceptual load of the stimuli. The response times for the unprocessed and CVR modified stimuli for VCV utterances at different SNRs for the individual subjects along with means and standard deviations are listed in Table 4.5. As seen in Figure 4.12, the mean response time increased as the SNR decreased from

Table 4.5 CVR Modification (Exp. I): Response times (s) for VCV utterances (unp: unprocessed stimuli, cvr: CVR modified). *p*: significance level of one-tailed paired t-test.

Subj.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	2.89	3.11	3.16	3.25	3.21	3.33	3.18	3.91	3.67	3.64	3.44	3.18
SC	2.79	2.75	3.91	3.79	3.84	3.94	4.25	3.86	4.04	3.9	4.10	4.11
SK	3.37	3.40	3.93	3.70	3.39	3.72	3.86	3.72	3.54	3.57	3.66	3.46
SB	3.43	3.19	3.25	3.01	2.96	3.00	2.92	2.94	3.22	2.84	3.18	3.27
SN	3.37	3.40	3.93	3.71	4.05	5.37	4.54	3.95	3.5	3.57	3.66	3.46
Avg.	3.17	3.17	3.64	3.49	3.49	3.87	3.75	3.68	3.59	3.50	3.61	3.5
s.d.	0.30	0.27	0.39	0.34	0.45	0.91	0.69	0.42	0.30	0.40	0.34	0.36
Impr.		0		-0.15		0.38		-0.07		-0.09		-0.11
p						n.s.						

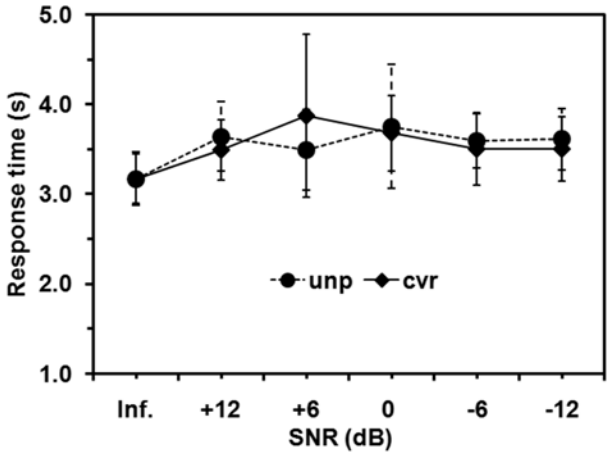


Figure 4.12 CVR Modification (Exp. I): Response times (s) for CVR modification *vs* SNR (dB). Error bars indicate standard deviations.

no-noise condition to +12 dB and there was no significant change for further decrease in SNR. There were no consistent differences in the values of the response time for the unprocessed and the CVR modified stimuli, indicating that processing for CVR modification resulted in an increase in recognition scores without a significant change in the perceptual load.

4.6 Results for time-scale modification (Exp. II)

The recognition scores for the individual subjects for the unprocessed (unp), synthesized without any modification (syn), and time-scale modified (tsc) stimuli along with the mean and standard deviation are given in Table 4.6. A plot of the mean scores is shown in Figure 4.13. The recognition scores of the unprocessed (unp) and the synthesized (syn) stimuli were

Table 4.6 Time-scale modification (Exp. II): Recognition scores (%) of VCV utterances (unp: unprocessed, syn: synthesized, tsc: time-scale modified). *p*: significance level of one-tailed paired t-test.

Subj.	SNR (dB)								
	∞			12			6		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
SA	100	98	96	96	95	96	93	91	88
SB	100	100	97	96	94	98	93	89	92
SC	100	100	100	99	99	99	91	92	91
SD	100	100	100	96	95	96	95	95	92
SE	100	98	98	96	96	96	95	91	87
Avg.	100	99	98	97	96	97	93	92	90
s.d.	0	1.1	1.8	1.3	1.9	1.4	1.7	2.2	2.3
Impr.		−1	−2		−1	0		−1	−3
p									

Subj.	SNR (dB)								
	0			−6			−12		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
SA	79	83	83	57	60	76	35	37	47
SB	79	81	84	57	58	69	36	36	47
SC	83	85	88	78	73	76	55	46	53
SD	84	85	92	71	73	73	47	47	53
SE	85	84	82	65	65	71	45	47	55
Avg.	82	84	86	66	66	73	44	43	51
s.d.	2.8	1.7	4.1	9.1	7.0	3.1	8.3	5.6	3.7
Impr.		2	4		0	7		−2	7
p			n.s.			n.s.			n.s.

Table 4.7 Time-scale modification (Exp. II): Recognition scores (%), averaged across subjects, for different vowel contexts.

Vow. cont.	SNR (dB)								
	∞			12			6		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
a	100	99	98	97	96	97	94	91	91
i	99	98	98	98	96	97	94	92	91
u	100	100	98	96	97	97	93	91	91

Vow. cont.	SNR (dB)								
	0			−6			−12		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
a	83	85	85	66	66	76	44	42	53
i	83	82	87	65	67	72	46	46	53
u	81	85	86	67	67	75	44	43	50

almost the same indicating no loss of intelligibility during synthesis. For SNRs of 0, –6, and –12 dB, the recognition scores of the unprocessed stimuli were 82, 66, and 44%, respectively. The corresponding values for the time-scale modified stimuli (tsc) were 86, 73, and 51%. Time-scale modification improved the recognition scores by nearly 4, 7, and 7% at SNRs of

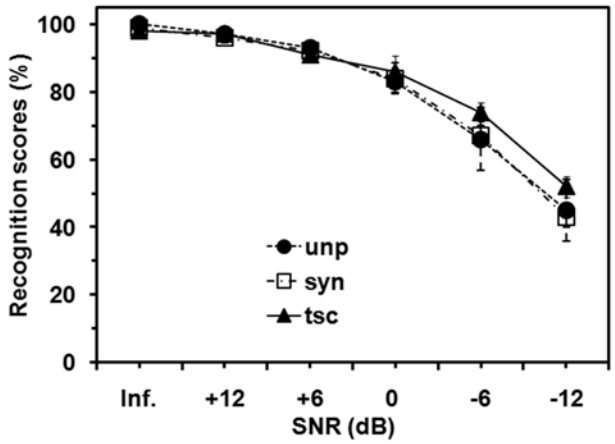


Figure 4.13 Time-scale Modification (Exp. II): Recognition scores (%) averaged *vs* SNR (dB) for VCV utterances. Error bars indicate standard deviations.

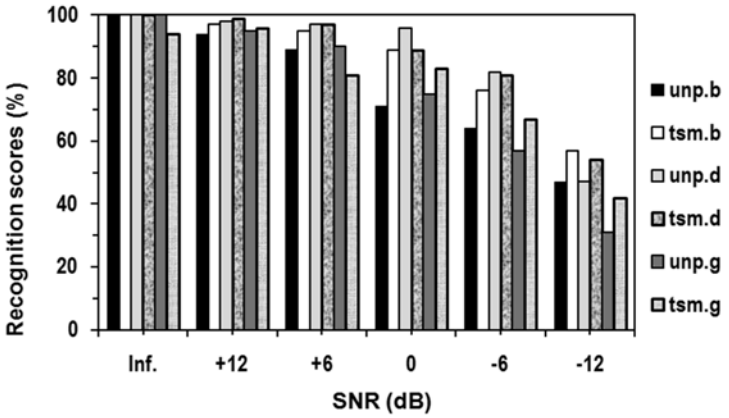


Figure 4.14 Time-scale Modification (Exp. II): Recognition scores (%) for voiced stops.

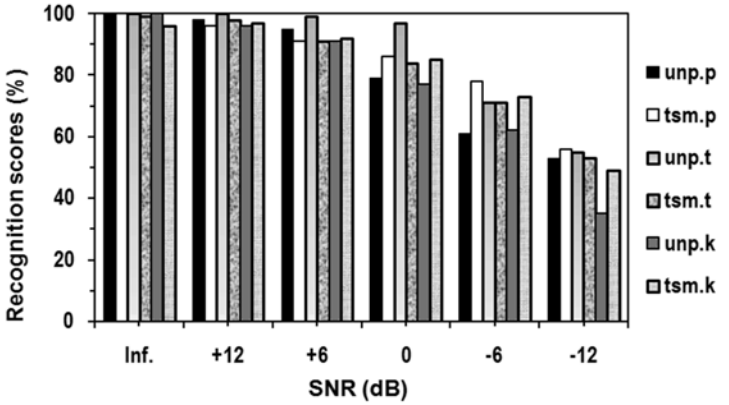


Figure 4.15 Time-scale Modification (Exp. II): Recognition scores (%) for unvoiced stops.

Table 4.8 Time-scale modification (Exp. II): Recognition scores (%) of individual stop consonants.

Stop	SNR (dB)								
	∞			12			6		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
b	100	100	100	94	97	97	89	95	95
d	100	100	100	98	98	99	97	97	97
g	100	96	94	95	94	96	90	89	81
p	100	100	100	98	95	96	95	89	91
t	100	100	99	100	98	98	99	94	91
k	100	98	96	96	95	97	91	86	92

Stop	SNR (dB)								
	0			−6			−12		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
b	71	86	89	64	68	76	47	53	57
d	96	90	89	82	78	81	47	38	54
g	75	79	83	57	56	67	31	35	42
p	79	82	86	61	67	78	53	48	56
t	97	87	84	71	68	71	55	50	53
k	77	80	85	62	64	73	35	37	49

Table 4.9 Time-scale modification (Exp. II): Relative information transmission (%). unp.: unprocessed stimuli, syn.: synthesized stimuli, tsc: time-scale modified stimuli.

SNR (dB)	overall			place			voicing		
	unp	syn	tsc	unp	syn	tsc	unp	syn	tsc
∞	99	97	96	99	96	93	99	99	100
12	92	92	90	92	87	90	90	90	92
6	84	84	80	81	81	72	88	82	87
0	66	67	71	57	56	59	73	76	88
–6	43	47	55	27	25	36	57	65	79
–12	22	23	28	7	5	11	35	38	46

(a): Overall (b): Place (c): Voicing

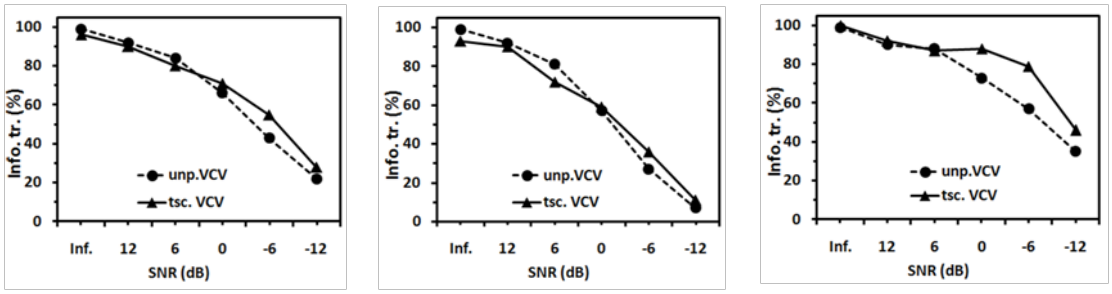


Figure 4.16 Time-scale Modification (Exp. II): Relative information transmission (%).

Table 4.10 Time-scale modification (Exp. II): Response times (s) for VCV utterances (unp: unprocessed stimuli, tsc: time-scale modified). *p*: significance level of one-tailed paired t-test.

Subj.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	tsc	unp	tsc	unp	tsc	unp	tsc	unp	tsc	unp	tsc
SA	2.89	2.69	3.16	3.01	3.21	3.29	3.18	3.27	3.67	3.08	3.44	2.86
SB	2.92	2.75	3.42	2.98	3.21	3.02	3.28	2.92	3.67	2.87	3.44	3.46
SC	3.06	3.61	3.91	3.34	3.84	3.14	4.25	3.52	4.04	3.14	4.10	3.81
SD	3.20	3.29	3.93	3.93	3.39	3.39	3.86	3.30	3.52	3.22	3.66	3.66
SE	3.17	3.01	3.66	3.41	3.43	3.19	3.78	3.24	3.50	3.20	3.67	3.50
Avg.	3.04	3.07	3.61	3.33	3.42	3.20	3.67	3.25	3.68	3.10	3.66	3.46
s.d.	0.14	0.38	0.33	0.38	0.26	0.14	0.44	0.22	0.22	0.14	0.27	0.37
Impr.		0.02		-0.28		-0.21		-0.42		-0.58		-0.20
p		n.s.										

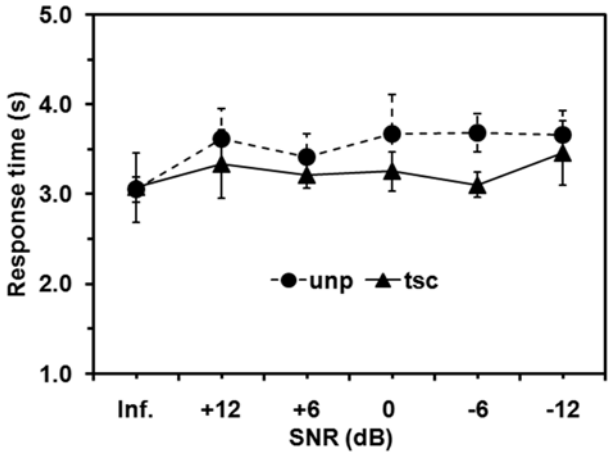


Figure 4.17 Time-scale Modification (Exp. II): Response times (s) for time-scale modification *vs* SNR (dB). Error bars indicate standard deviations.

0, –6, and −12 dB. The improvements were not statistically significant. The increase in recognition scores averaged across the subjects corresponded to an SNR advantage of 2 dB (measured using the method as described earlier).

Recognition scores averaged across subjects for the three vowel contexts are listed in Table 4.7 and these show no effect of the vowel context. Consonant-wise analysis of the recognition scores was performed to investigate the effect of time-scale modification on the recognition of individual stop consonants and the values are listed in Table 4.8. Recognition scores for stop consonants grouped into voiced and unvoiced at different SNRs are shown in Figures 4.14 and 4.15, respectively. At lower SNRs (< 0 dB), the labial (/b, p/) and velar (/g,

k/) stop consonants were more benefitted by time-scale modification than the alveolar stops (/d, t/).

Information transmission analysis (Miller and Nicely, 1955) was carried out on the stimulus-response confusion matrices to get the relative information transmitted in terms of overall, voicing, and place features. The results for overall, place, and voicing features are plotted in Figure 4.16 and listed in Table 4.9. The results show that the loss in information at the lower SNRs was mainly due to the place feature. In contrast to CVR modification which was effective in improving transmission of place feature, processing by time-scale modification of transition segments was more effective in improving the transmission of the voicing feature. The response times for the unprocessed and time-scale modified stimuli for VCV utterances at different SNRs for the individual subjects along with means and standard deviations are listed in Table 4.10. A plot of response times at different SNRs is shown in Figure 4.17. The mean response time increases as the SNR decreases to 12 dB and there is no significant change for further decrease in SNR. The response times for the time-scale modified (tsc) stimuli are lower than that of the unprocessed (unp) stimuli at all SNR levels below 12 dB but the improvements are not statistically significant.

4.7 Discussion

Modification of speech characteristics using acoustic properties of clear speech was investigated. Nonsense VCV utterances involving 6 stop consonants (/b, d, g, p, t, k/) in 3 vowel contexts (/a, i, u/) recorded from 5 speakers was used as the test material.

CVR modification was performed by 9 dB amplification of the VC and CV transition segments. The regions for modifications were selected using an automated landmark detection method. Time-scale modification was performed using sinusoidal model based analysis/synthesis. It involved expansion of the CV transition by a factor of 1.5. Results of the experiment with CVR modification (Exp. I) indicated improvements in recognition scores by nearly 7, 18, and 25% at SNRs of 0, −6, and −12 dB. The improvements were statistically significant at lower SNRs. The improvement in recognition scores was equivalent to an SNR advantage of 6 dB. CVR modification was equally effective in all the vowel contexts. It was effective in improving the transmission of place feature at lower SNRs. Consonant-wise analysis of results indicated CVR modification to be most effective for improving recognition of alveolar stops (/d, t/) compared to labial (/b, p/) and velar (/g, k/) stops. The response time was nearly the same for the unprocessed and CVR modified stimuli indicating no significant increase in the perceptual load introduced by the processing for CVR modification.

In the experiment with time-scale modification (Exp. II), the recognition scores for the synthesized stimuli were almost the same as those for the unprocessed stimuli at all SNRs, indicating no loss of intelligibility during synthesis using the sinusoidal model. Increase in the

transition durations did not result in any audible distortions. Time-scale modification improved the recognition scores by 4, 7, and 7% at SNRs of 0, −6, and −12 dB, respectively. The improvements were not statistically significant. The improvement in recognition scores was equivalent to an SNR advantage of 2 dB. No significant differences were observed between the recognition scores in the three vowel contexts. It was effective in improving the transmission of voicing feature at lower SNRs. Time-scale modification improved recognition of labial (/b, p/) and velar stops (/g, k/) as compared to alveolar (/d, t/) stops.

The results of the two experiments show that CVR modification is more effective than time-scale modification in improving recognition of stop consonants at lower SNRs. The contribution of the time-scale modification towards speech perception appears to be complementary to that of the CVR modification.

Chapter 5

REAL-TIME SPEECH INTELLIGIBILITY ENHANCEMENT USING CVR MODIFICATION

5.1 Introduction

Investigations presented in the previous chapter have shown automated detection and modification of acoustically abrupt landmarks to be effective in improving speech intelligibility. In comparison to time-scale modification, CVR modification is more effective in improving perception of consonants and it is computationally less expensive. This chapter presents a signal processing technique for CVR modification which is well suited for real-time implementation. It has been evaluated by conducting listening tests on normal hearing subjects with speech-shaped noise at different SNRs as a masker. The technique has been implemented for real-time operation using a DSP board based on a 16-bit fixed point processor with on-chip FFT hardware.

5.2 Signal processing for CVR modification

The regions for modification are located using the method SMTA described earlier in Section 3.5. The modification is carried out by applying the selected gain on the signal samples. The signal processing technique is shown as a block diagram in Figure 5.1. Processing is carried out using windowed frames of length 6 ms with a frame shift of 1 ms. The gain to be applied at frame position *n* is calculated using three inputs: first difference of the spectral centroid $dF_c(n)$, smoothened short-time energy $E_s(n)$, and peak energy $E_p(n)$. The frame energy $E(n)$ is smoothened by a 20-point moving average filter to get $E_s(n)$ as an indicator of energy variations at the phoneme level. The envelope of frame energy $E(n)$ is tracked using a peak detector with exponential decay to get the peak energy $E_p(n)$ as the following

$$E_p(n) = \begin{cases} E(n), & E(n) \geq E_p(n-1) \\ \alpha E_p(n-1), & \text{otherwise} \end{cases} \tag{5.1}$$

Use of $\alpha = 0.5^{1/200}$, with frame shift of 1 ms, corresponds to half-value release time of 200 ms, and the resulting $E_p(n)$ tracks the vowel energy and retains it during stop closures and other low energy clusters. The spectral centroid is computed by adding a 100 Hz tone with a level of −20 dB with reference to the peak energy. The magnitude spectra of 6 ms frames, calculated using Hanning window and 512-point FFT, are smoothened by *M*-frame moving average. From the smoothened spectrum, the spectral centroid $F_c(n)$ is computed and is smoothened by *L*-point median filtering for

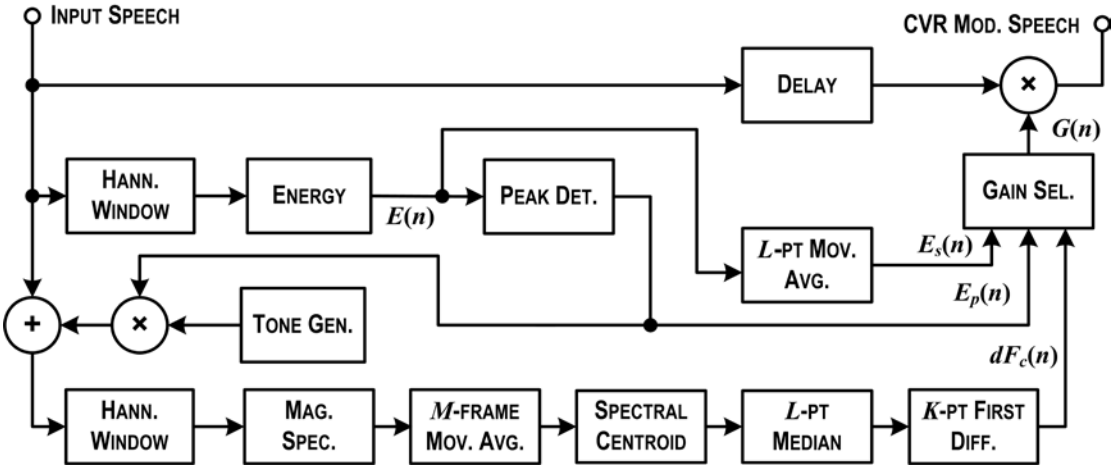


Figure 5.1 Signal processing for CVR modification.

suppressing ripples without significantly smearing the changes due to major spectral transitions. For detecting changes in the spectral centroid, its *K*-point first difference $dF_c(n)$ is computed using Eq. (3.27). Use of *K*, *L*, *M* corresponding to 20 ms time step was found to be optimal for detecting spectral transitions.

The gain selection for CVR modification uses a hysteresis based thresholding of $dF_c(n)$ with upper and lower thresholds of θ_h and θ_l. Hysteresis based thresholding is used to reduce the momentary triggering of an unintended action caused by random parameter variations around the threshold values. It uses an upper and a lower threshold. The action is taken when the parameter exceeds the upper threshold and it is reset when it falls below the lower threshold. In our application, a threshold value is used on the rate of change of centroid to trigger the gain applied for CVR modification. A hysteresis based thresholding ensures that random fluctuations in the rate of change of centroid (just above the lower threshold) do not trigger CVR modification and momentary dips (just below the upper threshold) do not reset the gain applied for CVR modification. It is carried out with the help of a flag updated at each frame position as

$$\begin{aligned} \mathrm{CVR}(n) = 1, &\quad dF_c(n) > \theta_h \\ 0, &\quad dF_c(n) < \theta_l \\ \mathrm{CVR}(n-1), &\quad \theta_l \le dF_c(n) \le \theta_h \end{aligned} \tag{5.2}$$

The threshold values of 350 Hz and 300 Hz were selected as θ_h and θ_l, respectively, based on observations of the plots of the first difference of spectral centroid for the keywords in the speech material recorded with MRT wordlist. Hysteresis based thresholding with these values prevents momentary fluctuations in $dF_c(n)$ from triggering CVR modification, without missing actual abrupt spectral transitions. The maximum gain for enhancing a segment is set as A_m subjected to a condition

that the energy of the frame after its amplification does not exceed the peak energy. The maximum gain for a frame is calculated as

$$G_m(n) = \min[A_m, (E_p(n)/E_s(n))^{1/2}] \tag{5.3}$$

To avoid perceptible distortions caused by abrupt changes, the gain is changed from its current value to the target value in *p* logarithmic steps of

$$\gamma = [G_m(n)]^{1/p} \tag{5.4}$$

The gain to be applied for a frame *n* is calculated as

$$G(n) = \begin{cases} \min[G(n-1)\gamma,\ G_{\mathrm{m}}(n)], & \mathrm{CVR}(n) = 1 \\ \max[G(n-1)/\gamma,\ 1], & \text{otherwise} \end{cases} \tag{5.5}$$

To provide significant enhancement of the transition segments without introducing perceptible distortions, we have used a maximum gain of 9 dB (i.e. A_m = 2.82) and p = 3. The input signal is multiplied by the gain to get the CVR modified signal. A delay of 10 ms is introduced in the signal path to approximately compensate for the delay in the detection of spectral transitions due to the averaging of the magnitude spectrum and the 20-point median filter.

Examples for CVR modification performed on utterances "*you will mark ut please*", and "*would you write tick*" are shown in Figures 5.2 and 5.3, respectively. The figures show the waveforms and spectrograms of the original and CVR modified utterances along with the centroids, its first differences, and the gains for amplifying the stop release bursts and frications in the utterances. Examples of the speech files used as the test material are available on the web (Jayan, 2014).

5.3 Listening tests

Effectiveness of the processing in improving consonant recognition under adverse listening conditions was evaluated by conducting two experiments involving listening tests on normal-hearing subjects for closed-set recognition of consonants in the presence of noise. The first set of tests, named as Experiment III, was conducted using nonsense CV and VC syllables with six stop consonants and three vowels as the test material. Use of this material helped in evaluating the recognition of individual stop consonants in the two positions separately and the contribution of different features in the improvement of recognition scores. The second set of tests, named as Experiment IV, was conducted with MRT wordlist consisting of rhyming CVC words. The wordlist has different types of consonants (stops, fricatives, etc) in initial and final positions and different vowels. It has a better representation of the occurrence of consonants in speech than the nonsense CV and VC syllables. It

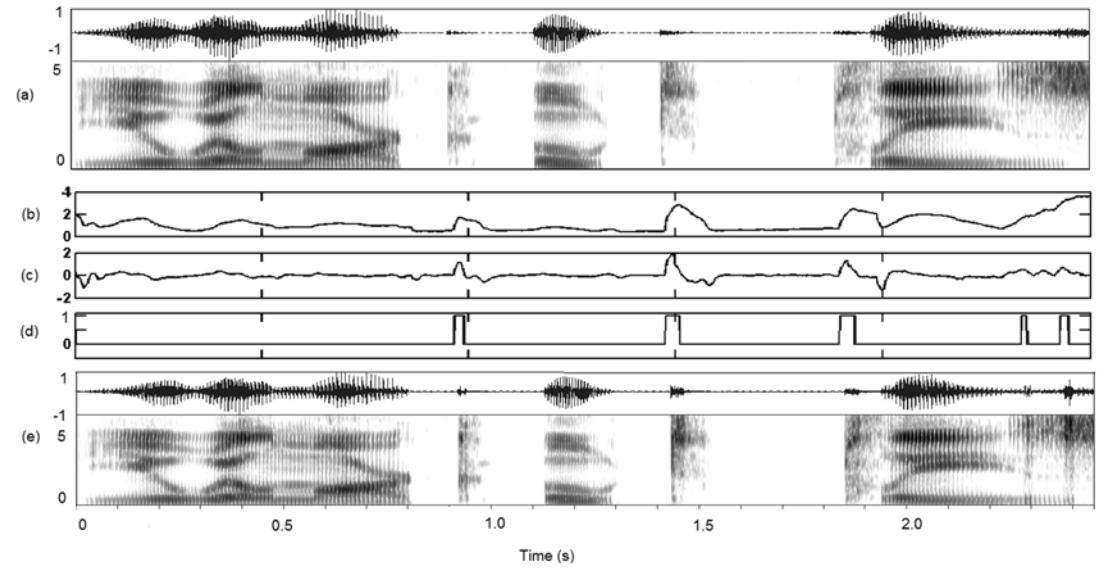


Figure 5.2 Example of CVR modification: (a) speech signal of the utterance "*you will mark ut please*" and its spectrogram, (b) spectral centroid $F_c(n)$, (c) first difference of centroid $dF_c(n)$, (d) windows selected for CVR modification, (e) CVR modified signal and its spectrogram. Frequency axis of spectrogram in kHz.

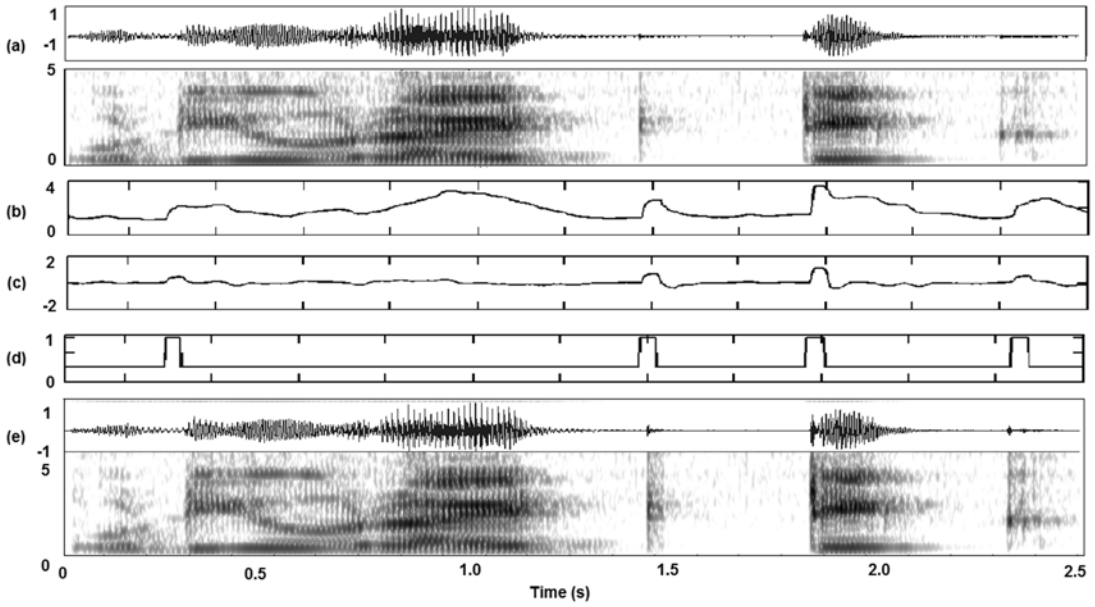


Figure 5.3 Example of CVR modification: (a) speech signal of the utterance "*would you write tick*" and its spectrogram, (b) spectral centroid $F_c(n)$, (c) first difference of centroid $dF_c(n)$, (d) windows selected for CVR modification, (e) CVR modified signal and its spectrogram. Frequency axis of spectrogram in kHz.

gives a recognition score averaged over vowel contexts, consonant positions, and different consonants. As it has meaningful words, it requires minimal training for listeners (House et al, 1965; ANSI, 1989).

The stimuli were processed for 9 dB CVR modification, using the method as described earlier. Speech-spectrum shaped noise with spectral envelope similar to the long-term spectral

envelope of speech was subsequently added as a masker. Its spectrum was flat from 100 Hz to 1 kHz and had a slope of −12 dB/octave thereafter. The noise level was constant over the length of the speech stimulus, with the SNR calculated with respect to the RMS value of the signal frame with the highest energy. The noise extended by 1 s on either side of the stimulus. The tests were conducted for SNR values between −12 dB and 12 dB, because the repeatability of the responses for SNR below −12 dB was very low and the masking effect of noise for SNR above 12 dB was found to be negligible. As in the case of Experiment I and Experiment II, the response choice and the response time taken by the subject to mark the response were recorded.

5.3.1 Tests with CV and VC syllables (Experiment III)

The test material for Experiment III consisted of nonsense CV and VC syllables with consonants /b, d, g, p, t, k/ paired with vowels /a, i, u/ as keywords in the carrier sentence "You will mark −−−− please". There were a total of 36 utterances, 18 with VC syllables and 18 with CV syllables. The test material was recorded from a male speaker in an acoustically treated room, using B&K microphone model 2210, with sampling frequency of 10 kHz and 16-bit quantization. All the speech stimuli were normalized to have the same RMS value. To reduce the effect of silence segments on the RMS value, the input stimulus was segmented in 20 ms frames and only the frames with energy within 20 dB of the frame with the highest energy were used for calculating the RMS value. The signals were processed for 9 dB CVR modification. The test stimuli were generated by adding noise at SNRs of ∞ (quiet), 6, 3, 0, −3, −6, −9, and −12 dB.

Listening tests were conducted using a computerized test administration setup with a graphical user interface in an acoustically treated room. The stimuli were presented through Sennheiser PX80 headphones and the presentation level was set at the most comfortable level selected by the subject and the same level was maintained across all listening conditions. The subject clicked on a 'play' button to listen to the stimulus and responded by clicking on one of the six response choices displayed on the computer screen. The response choices were the six possible syllables with the same vowel. The effect of position bias was eliminated by randomizing the position of the response choices. After each presentation, the response and the response time were recorded. Presentation-response process was repeated for all presentations in a test session. The subjects were given training with noise-free unprocessed stimuli to familiarize them with the stimuli and the test process.

Ten normal-hearing subjects (6 male and 4 female, age: 21−45 years) participated in the listening tests. With two types of stimuli (unprocessed and processed) and eight SNRs, there were 16 listening conditions. For each listening condition, each of the 36 stimuli was presented five times. These 180 presentations were distributed in a randomized order over six sessions, each with 30 presentations and taking about 5−10 minutes for completion. A subject participated in a maximum of six test sessions, with breaks in between, on a day. The listening tests for each subject took about

16–20 hours, distributed over about 30 days. In order to reduce response biases due to practice or fatigue, the order of the listening conditions was randomized across the subjects. The instructions given to the subjects and the forms for collection for background information of the subjects and their consent to participate in the tests are given in Appendix A.

The responses to the stimuli were tabulated as stimulus-response confusion matrices and these were used to obtain recognition scores as a function of SNR. The differences between the recognition scores for the unprocessed speech and those for the processed speech are indicators of enhancement in intelligibility due to processing at different SNRs. The SNR advantage because of processing was measured using the method as described earlier in Section 4.4. The stimulus-response confusion matrices were also used to carry out information transmission analysis (Miller and Nicely, 1955) to get the relative information transmitted in terms of overall, voicing, and place features.

5.3.2 Tests with MRT wordlist (Experiment IV)

The test material for Experiment IV consisted of MRT wordlist with monosyllabic words of CVC form. The list had 50 sets, each set having six words with the same vowel in the middle. Either initial or final consonant remained the same, while the other consonant was different in each word. Each word was recorded as a keyword in the carrier sentence "Would you write ––––". The 300 words (i.e. 50 sets × 6 words in each set) were arranged in 6 test lists (1x, 1y, 2x, 2y, 3x, 3y) of 50 words each, using a two-level randomization process with the set level (1, 2, 3) and the word level (x, y). The recording and preparation of the stimuli followed the same process as in Experiment III, with SNR of ∞, 12, 6, 0, –6, and –12 dB.

The test administration setup and the testing process was the same as for Experiment III, with the six rhyming words corresponding to the stimulus displayed as the response choices. Ten normal-hearing subjects (6 male and 4 female, age: 18–40 years) participated in the tests. The experiment involved 12 listening conditions, with two types of stimuli (unprocessed and processed) and six SNRs. Each test list with 50 stimuli was presented in one session, taking 15-20 minutes for completion. On a day, a subject participated in a maximum of three test sessions, with breaks in between. The tests were spread across a period of one month, each subject spending nearly 18–24 hours. As in the first experiment, the order of the listening conditions was randomized across the subjects in order to reduce the effect of practice and fatigue. The instructions given to the subjects and the forms for collection for background information of the subjects and their consent to participate in the tests are given in Appendix A. The word list used for MRT is given in Appendix B.

The responses to the stimuli were used to obtain recognition scores as a function of SNR. The differences between the recognition scores for the unprocessed speech and those for the processed speech are indicators of enhancement in intelligibility due to processing at different SNRs. The SNR advantage because of processing was measured using the method as described earlier in Section 4.4.

5.3.3 Results of Experiment III: Listening tests with CV and VC syllables

The recognition scores for the unprocessed and the processed stimuli for CV and VC syllables at different SNR, for the individual subjects along with the means and standard deviations, are given in Table 5.1. It also gives the improvement in the scores after processing and significance for paired t-test. For the unprocessed stimuli, the recognition scores of all the subjects decreased with decrease in SNR, from 100% at no noise to approximately 50% at −12 dB SNR. The scores after CVR modification increased for all the subjects. The improvements were higher at lower SNRs and they were statistically significant ($p < 0.001$). The mean recognition scores are plotted in Fig. 5.4. The mean improvements in scores after CVR modification for CV syllables were 8, 8, 9, and 19% at SNR of 6, 0, −6, and −12 dB, respectively. The corresponding improvements for VC syllables were 7, 9, 11, and 14%. The increase in recognition scores averaged across the subjects corresponded to an SNR advantage of 6 dB for CV syllables and 5 dB for VC syllables.

For the unprocessed stimuli at negative SNRs, the scores for the VC syllables were generally lower than those for CV syllables. This could be due to masking of the low energy consonant segment by the preceding high energy vowel segment. After processing, the recognition scores for the two sets of syllables were generally the same, indicating the effectiveness of CVR modification in reducing forward masking.

The perception of a stop consonant in noise depends on the robustness of acoustic cues distributed in the intensity, spectral, and temporal domains. The pattern of consonant confusions in noise depends on the type of the consonant, vowel context, and the properties of noise masker (Regnier and Allen, 2008). The recognition scores for individual stop consonants and the scores in the three vowel contexts were also analyzed. The recognition scores averaged across subjects for the individual stop consonants for CV and VC syllables are given in Table 5.2. The corresponding plots are given in Figure 5.5. The improvements for the alveolar stops (/t, d/) were higher than the improvements observed for labial (/p, b/) and velar (/k, g/) stops. The averaged recognition scores in the three vowel contexts for CV and VC syllables are given in Table 5.3. The corresponding plots are given in Figure 5.6. The scores for the CVR modified stimuli were generally higher than those of the unprocessed stimuli irrespective of the vowel contexts. The perception of consonants in vowel context /u/ was found to be poorer than that of vowel contexts /a/ and /i/ for both CV and VC syllables. CVR modification was most effective in improving consonant recognition scores in the context of vowel /u/. Both these observations are in agreement with the results reported for CV syllables by Hazan and Simpson (1998).

The confusion matrices at different SNRs for CV and VC syllables were analyzed. For CV syllables, the lowest recognition scores were obtained for unvoiced velar stop /k/ at lower SNRs. This was mainly due to the confusion for the stimulus-response pair (/k, p/) in the vowel context /a/. CVR modification was effective in reducing the (/k, p/) confusions for SNRs ranging from −6 to +6 dB. The major confusions in the /i/ and /u/ vowel contexts were for the stimulus-response pair (/p, k/).

Table 5.1 Experiment III: Recognition scores (%) at different SNRs for listening tests with nonsense CV and VC syllables. unp: unprocessed stimuli, cvr: stimuli processed with CVR modification. *p*: one-tailed significance level of paired t-test (n = 10, df = 9).

a) Recognition scores for CV syllables

Subj.	SNR (dB)															
	∞		6		3		0		−3		−6		−9		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	100	100	86	100	84	94	78	94	72	83	60	78	56	68	44	66
SB	100	100	91	100	90	98	88	91	83	86	76	80	61	77	48	72
SC	100	100	87	96	81	83	76	83	68	79	71	70	61	69	47	62
SD	100	100	90	100	91	96	87	93	77	86	70	80	66	78	53	76
SE	100	100	90	100	91	99	90	93	84	88	74	77	69	78	60	73
SF	100	100	93	96	91	94	88	92	82	91	81	89	61	88	47	76
SG	100	100	91	100	90	99	87	99	72	99	70	82	63	81	51	68
SH	100	100	97	100	90	98	89	99	74	99	62	81	63	81	54	68
SI	100	100	87	91	90	90	86	94	76	93	78	83	63	82	51	69
SJ	100	100	90	100	90	99	88	99	73	99	71	83	64	81	51	69
Mean	100	100	90	98	89	95	86	94	76	90	71	80	62	78	51	70
s.d.	0	0	3.2	3.1	3.4	5.2	4.7	4.8	5.4	7.1	6.5	4.9	3.4	6.0	4.5	4.4
Impr.				8.1		6.3		8.0		14.2		9.0		15.6		19.3
s.d.				3.6		3.4		4.3		9.0		6.5		5.6		5.1
t				7.1		5.7		5.9		4.9		4.4		8.8		12.0
p				<0.001		<0.001		<0.001		<0.001		<0.001		<0.001		<0.001

b) Recognition scores for VC syllables

Subj.	SNR (dB)															
	∞		6		3		0		−3		−6		−9		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	100	100	92	99	88	89	79	89	70	87	60	79	58	76	41	61
SB	100	100	91	100	91	96	87	89	82	89	71	83	52	74	43	67
SC	100	100	94	94	93	93	82	97	78	86	72	78	61	77	61	63
SD	100	100	90	94	89	91	79	89	74	83	67	82	56	71	51	68
SE	100	100	90	94	89	91	80	89	76	87	60	78	59	76	54	67
SF	100	100	89	91	88	84	82	82	76	86	73	81	53	76	61	68
SG	100	100	82	100	80	94	77	89	76	80	72	78	59	73	51	66
SH	100	100	89	100	83	94	79	89	76	81	67	78	60	73	49	64
SI	100	100	93	100	86	96	78	90	81	87	71	81	59	74	51	66
SJ	100	100	86	100	82	94	77	89	76	83	72	78	62	73	51	63
Mean	100	100	90	97	87	92	80	89	77	85	69	80	58	74	51	65
s.d.	0	0	3.5	3.5	4.1	3.6	3.0	3.6	3.4	3.0	4.9	2.0	3.3	1.9	6.5	2.4
Impr.				7.0		5.0		9.0		8.4		11.1		16.4		14.0
s.d.				5.6		6.0		4.6		3.7		4.8		3.7		6.2
t				4.3		2.7		6.2		7.1		7.2		13.7		7.1
p				<0.001		0.01		<0.001		<0.001		<0.001		<0.001		<0.001

CVR modification was effective in reducing these confusions for SNRs above 0 dB. Confusions were observed for the stimulus-response pairs (/b, d/) and (/t, p/) in the vowel context /i/ at SNRs below 0 dB. These confusions reduced the scores for unprocessed /b/ and /t/ at lower SNRs. CVR modification was effective in reducing these confusions. It was also noticed that for SNRs below −3 dB, CVR

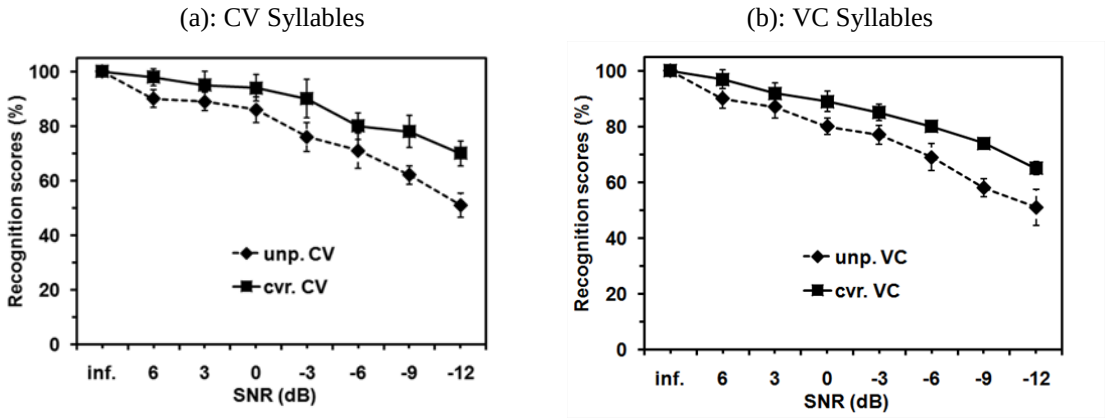


Figure 5.4 Experiment III: Recognition scores (%) averaged across subjects. Error bars indicate standard deviations.

Table 5.2 Experiment III: Recognition scores (%) at different SNRs for listening tests with nonsense CV and VC syllables for individual stop consonants. unp: unprocessed stimuli, cvr: stimuli processed with CVR modification.

a) Recognition scores for stops in CV syllables

Stop	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
b	99	100	96	95	97	95	90	96	78	91	65	81	55	81	62	74
d	100	99	99	100	100	100	100	100	98	100	88	100	79	99	59	89
g	100	100	98	100	100	99	97	99	90	89	78	76	81	66	49	50
p	100	100	76	98	69	86	78	88	82	89	83	84	77	88	74	85
t	100	100	98	99	99	100	92	98	71	100	71	100	41	100	37	98
k	100	100	73	97	69	91	55	83	39	72	43	41	46	35	23	23

b) Recognition scores for stops in VC syllables

Stop	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
b	99	100	98	100	93	96	89	99	85	93	88	92	83	91	73	81
d	100	100	99	100	100	100	79	100	71	100	51	100	41	97	50	63
g	100	100	70	90	61	85	54	69	56	63	47	49	26	23	13	15
p	100	100	98	96	96	97	95	97	93	97	91	95	83	91	85	91
t	100	100	100	100	99	98	100	100	94	100	81	100	75	100	58	100
k	100	100	73	98	71	78	63	70	60	55	54	41	40	43	29	40

modification increased (/k, t/) confusions in the vowel context /i/. These confusions reduced the effectiveness of CVR modification for unvoiced velar stop /k/ at lower SNRs. CVR modification increased the confusions for (/g, b/) in the context of vowel /u/ at SNRs below 0 dB. CVR modification was not very effective for velar stops /g/ and /k/ at lower SNRs. For VC syllables, lower recognition scores at lower SNRs were observed for consonants /g/ and /k/. This was mainly due to

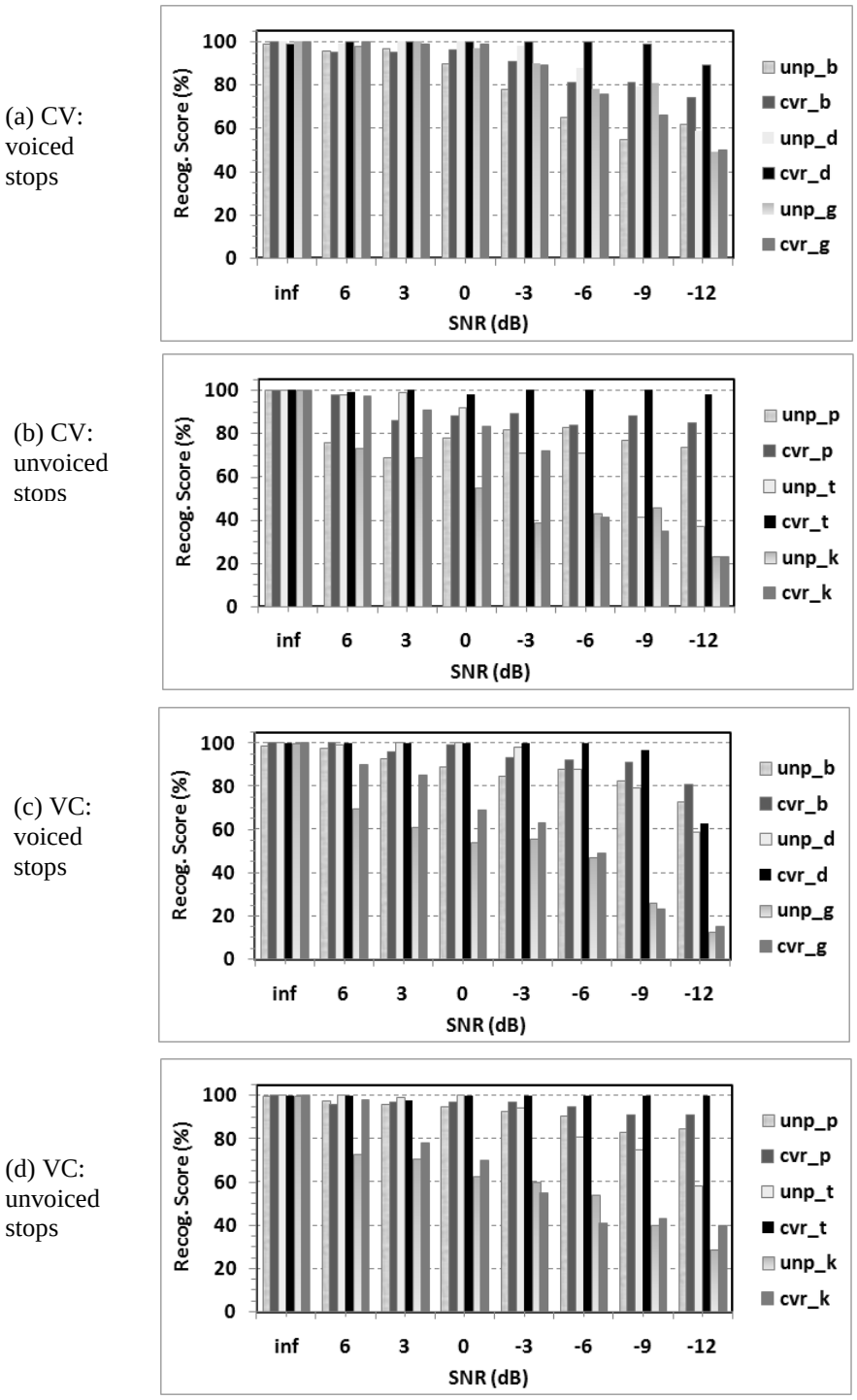


Figure 5.5 Experiment III: Recognition scores (%) for voiced and unvoiced stops in CV and VC syllables.

confusions for the stimulus-response pairs (/g, b/) and (/k, p/) in the context of vowel /u/. CVR modification was effective in reducing these confusions only at positive SNRs. Confusions also occurred for (/t, p/) in the vowel context /u/ at lower SNRs. CVR modification was effective in reducing these confusions.

Table 5.3 Experiment III: Recognition scores (%) at different SNRs for listening tests with nonsense CV and VC utterances for three vowel contexts. unp: unprocessed stimuli, cvr: stimuli processed with CVR modification.

a) Recognition scores for three vowel contexts in CV syllables

Vow.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
a	100	100	89	100	86	97	84	93	80	90	80	85	70	84	61	78
i	100	100	89	99	88	93	86	93	85	92	85	81	69	76	54	66
u	100	100	93	96	92	95	86	96	63	89	49	75	51	75	37	66

b) Recognition scores for three vowel contexts in VC syllables

Vow.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
a	100	100	98	94	94	97	94	98	92	92	85	84	70	71	52	58
i	100	100	96	100	95	100	80	100	77	95	72	89	68	87	64	77
u	100	100	75	98	71	80	66	69	60	67	49	66	36	65	38	61

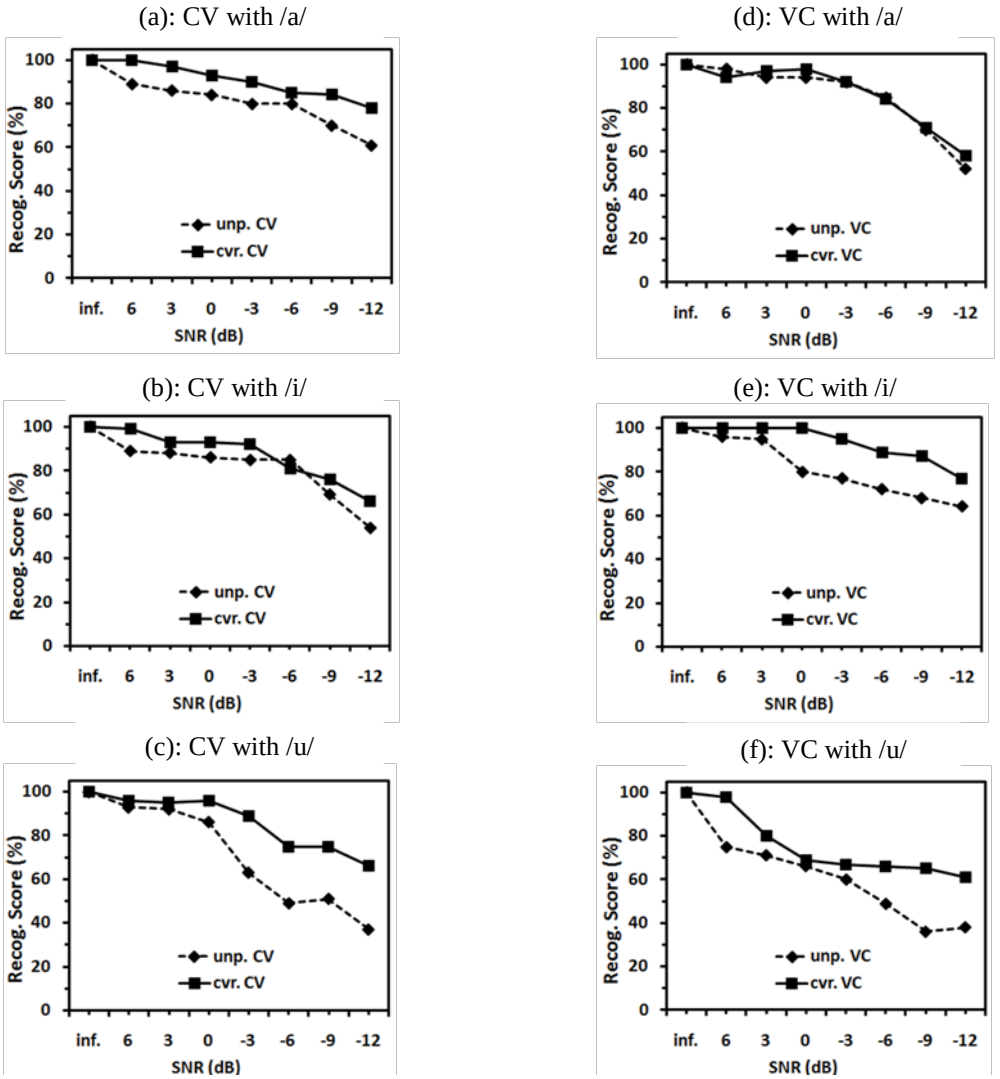


Figure 5.6 Experiment III: Recognition scores (%) averaged across subjects for unprocessed (dotted) and CVR modified (solid) for CV and VC syllables in three vowel contexts.

Table 5.4 Experiment III: Relative information transmission (%). unp: unprocessed stimuli, cvr: stimuli processed with CVR modification.

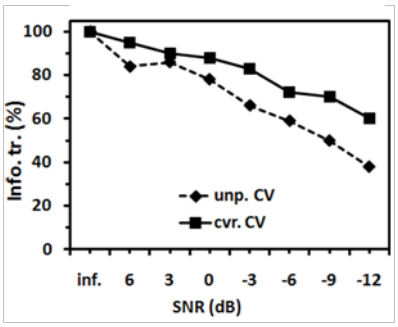
a) Relative information transmission (%) for CV syllables

Info. tr.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
Overall	100	100	84	95	86	90	78	88	66	83	59	72	50	70	38	60
Place	100	100	71	92	70	82	58	79	40	71	33	53	19	49	8	35
Voicing	100	100	96	100	100	99	99	99	99	100	99	99	94	100	80	94

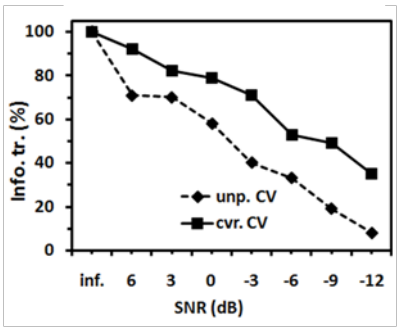
a) Relative information transmission (%) for VC syllables

Info. tr.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
Overall	100	100	84	94	81	85	70	84	66	80	57	75	49	69	36	54
Place	100	100	74	89	67	75	50	74	43	68	29	59	15	50	13	30
Voicing	100	100	99	98	100	100	98	98	100	100	97	97	93	98	68	79

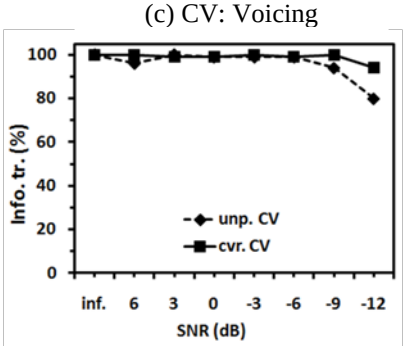
(a) CV: Overall



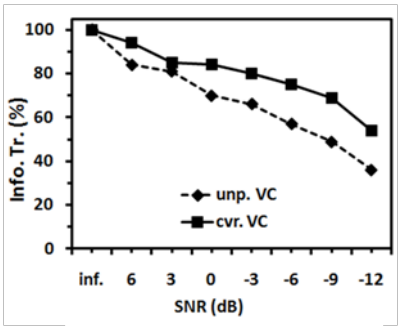
(b) CV: Place



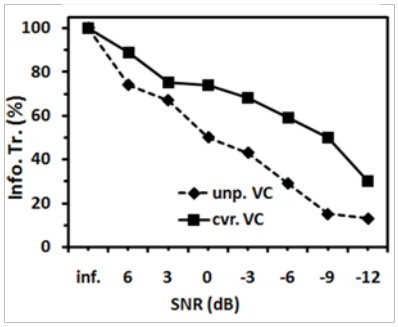
(c) CV: Voicing



(d) VC: Overall



(e) VC: Place



(f) VC: Voicing

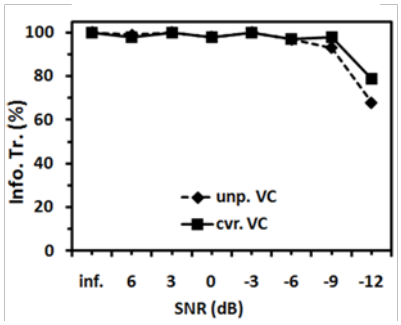


Figure 5.7 Experiment III: Relative information transmission for unprocessed (dotted) and CVR modified (solid) for CV and VC syllables.

Table 5.5 Experiment III: Response times (s) at different SNRs for listening tests with nonsense CV and VC syllables. unp: unprocessed stimuli, cvr: stimuli processed with CVR modification. *p*: one-tailed significance level of paired t-test (n = 10, df = 9).

a) Response time (s) for CV syllables

Subj.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	2.8	2.7	4.1	4.1	4.2	4.0	5.2	4.1	3.9	3.9	3.9	4.2	3.9	4.2	3.8	4.0
SB	2.7	2.6	3.9	3.8	3.8	3.7	3.9	4.0	4.4	4.5	4.5	4.6	4.5	4.6	4.3	4.6
SC	3.1	3.1	4.3	4.2	4.2	4.1	4.0	3.9	3.9	4.0	4.1	4.3	4.4	4.2	4.2	4.2
SD	2.7	2.6	2.9	3.9	3.6	4.0	5.0	3.7	4.3	3.8	4.2	3.8	3.7	3.8	3.8	3.8
SE	2.7	2.6	2.9	3.9	3.8	4.0	4.9	3.8	3.9	3.8	4.2	4.0	3.8	3.8	3.8	3.8
SF	4.4	3.5	5.1	4.5	4.4	4.5	4.6	4.3	4.5	4.2	4.4	4.5	4.6	4.5	4.2	3.8
SG	2.7	2.7	3.8	4.2	4.0	3.9	3.9	4.0	3.9	4.1	3.8	3.9	3.9	4.0	4.2	4.0
SH	2.7	2.7	3.9	4.2	3.9	3.9	3.8	3.9	3.9	4.0	4.0	3.9	3.9	4.0	4.2	4.0
SI	3.2	2.7	4.4	4.3	4.1	4.0	4.4	3.8	4.2	3.8	4.0	3.6	3.9	4.0	4.2	4.2
SJ	2.7	2.7	3.9	4.2	3.9	3.8	3.9	4.0	3.9	4.0	3.8	3.9	3.9	3.9	4.1	4.0
Mean	3.0	2.8	3.9	4.1	4.0	4.0	4.3	3.9	4.1	4.0	4.1	4.1	4.1	4.1	4.1	4.0
s.d.	0.5	0.2	0.6	0.2	0.2	0.2	0.5	0.1	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2
Impr.	–0.1		0.2				–0.4		–0.1						–0.1	
s.d.	0.3		0.5				0.5		0.2						0.2	
t			1.2													
p			n.s.													

b) Response time (s) for VC syllables

Subj.	SNR (dB)															
	∞		6		3		0		–3		–6		–9		–12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	2.9	2.8	4.2	3.9	4.4	4.0	6.0	4.6	3.9	4.0	4.0	3.8	3.9	4.2	3.9	4.3
SB	2.8	2.7	3.8	3.9	3.8	3.8	3.8	4.0	4.5	4.6	4.7	4.7	4.6	4.6	4.3	4.7
SC	3.3	3.3	4.3	4.3	4.2	4.1	4.2	4.0	4.0	4.1	4.2	4.4	4.3	4.2	4.2	4.3
SD	2.8	2.7	2.0	4.0	3.7	4.0	4.9	3.8	4.3	3.8	4.2	3.9	3.7	3.8	3.9	3.9
SE	2.8	2.7	2.0	3.8	3.7	3.9	4.8	3.8	4.0	3.8	4.1	4.1	3.8	3.8	3.8	4.0
SF	4.3	3.4	5.1	4.6	4.4	4.6	4.7	4.2	4.7	4.2	4.4	4.5	4.6	4.5	4.2	3.9
SG	2.8	2.7	3.8	4.1	3.9	3.9	3.9	4.0	3.9	4.2	3.9	4.0	3.9	4.0	4.2	4.0
SH	2.7	2.7	3.8	4.2	3.9	3.9	3.9	3.9	3.9	4.1	4.1	3.9	4.0	4.0	4.2	4.0
SI	3.3	2.8	4.3	4.2	4.2	4.0	4.4	3.8	4.3	3.8	4.0	3.8	3.9	4.0	4.2	4.2
SJ	2.6	2.7	3.8	4.1	3.9	3.8	3.9	3.9	3.9	4.1	3.9	3.9	3.8	3.9	4.3	3.9
Mean	3.0	2.8	3.7	4.1	4.0	4.0	4.5	4.0	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
s.d.	0.5	0.2	0.9	0.2	0.2	0.2	0.6	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2
Impr.	–0.2		0.4				–0.5									
s.d.	0.3		0.8				0.5									
t			1.5													
p			n.s.													

Information transmission analysis (Miller and Nicely, 1955) was carried out on the stimulus-response confusion matrices to get the relative information transmitted in terms of overall, voicing, and place features, for CV and VC syllables. The results listed in Table 5.4 and plotted in Figure 5.7 show that the loss in information at the lower SNRs was mainly due to place feature and the processing by CVR modification was effective in improving it. For CV syllables, the improvement in

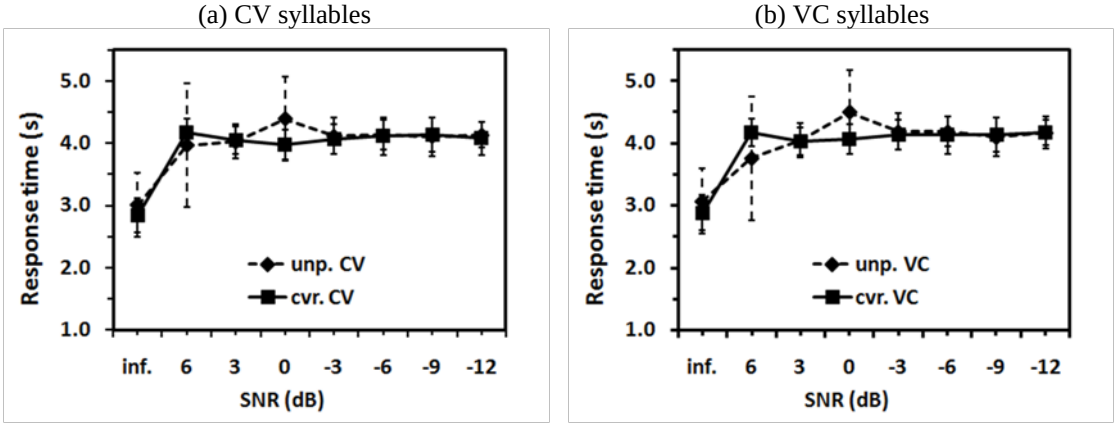


Figure 5.8 Experiment III: Response time (s) averaged across subjects. Error bars indicate standard deviations.

transmission of place feature was nearly 21, 20, and 27% at SNRs of 0, –6, and –12 dB, respectively. The corresponding improvements for VC syllables were by 24, 30, and 17%. The voicing feature was not much affected at SNRs ranging from 6 to –6 dB.

As in the case of Experiment I and II described in Chapter 4, analysis of response times were carried out for the unprocessed and CVR modified stimuli, to get an indication of the effect of CVR modification on the perceptual load. The response time averaged across subjects for CV and VC syllables are listed in Table 5.5 and also shown in Figure 5.8. The response time for all the subjects increased with decrease in SNR. There was no significant difference between the response time for the processed and the unprocessed stimuli, indicating that processing improved recognition scores without affecting the perceptual load.

5.3.4 Results of Experiment IV: Listening tests using MRT wordlists

The recognition scores for the individual subjects along with the mean and standard deviation are given in Table 5.6 and a plot of the mean scores is shown in Fig. 5.9. The recognition scores for the unprocessed stimuli decreased with decrease in SNR and the decrease was more than that observed for CV and VC syllables in Experiment III. The improvements gained by CVR modification were similar for all subjects. The mean improvements in the recognition scores due to CVR modification were 8, 9, and 11% at SNR of 0, −6, and −12 dB, respectively, and they were statistically significant ($p < 0.001$). The increase in recognition scores averaged across the subjects corresponded to an SNR advantage of 3 dB.

Analysis of recognition scores indicated the processing to be equally effective in improving recognition scores of consonants in word initial and word final positions. The responses for the individual consonants in the MRT wordlist were grouped, ignoring the position and vowel context, and the effect of CVR modification on recognition of each consonant was analyzed using the stimulus-response data for the ten subjects. It was found that processing improved the recognition of

Table 5.6 Experiment IV: Recognition scores (%) at different SNRs for listening test with MRT wordlist, unp: unprocessed stimuli, cvr: stimuli processed with CVR modification. *p*: significance level of one-tailed paired t-test.

Subj.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	96	97	93	95	90	92	77	85	59	67	26	36
SB	95	96	95	96	89	93	77	87	57	71	31	39
SC	96	97	93	95	85	88	76	84	54	64	26	36
SG	94	96	88	90	89	87	73	81	53	58	26	36
SH	95	97	93	96	91	89	78	87	58	65	28	38
SI	96	96	94	95	89	93	76	84	57	66	31	42
SJ	95	95	89	90	89	89	76	85	56	63	26	38
SK	96	97	94	94	90	93	78	85	58	67	26	33
SL	97	96	93	94	87	88	75	84	52	66	26	37
SM	96	97	93	95	89	90	75	82	56	65	30	40
Avg.	96	96	93	94	89	90	76	84	56	65	28	39
s.d.	0.7	0.6	2.1	2.3	1.5	2.1	1.5	1.8	2.4	3.2	2.1	1.9
Impr.		0		1		1		8		9		11
p		<0.05		<0.001		<0.05		<0.001		<0.001		<0.001

Table 5.7 Experiment IV: Response times (s) at different SNRs for listening tests with MRT wordlists unp: unprocessed stimuli, cvr: stimuli processed with CVR modification. *p*: one-tailed significance level of paired t-test (n = 10, df = 9).

Subj.	SNR (dB)											
	∞		12		6		0		−6		−12	
	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr	unp	cvr
SA	2.83	2.86	3.88	3.99	3.99	3.88	3.92	3.81	4.03	3.86	3.92	4.11
SB	2.78	2.68	3.88	3.72	3.95	3.86	3.78	3.95	4.09	4.00	3.99	4.09
SC	2.83	2.86	3.88	3.99	4.3	4.37	4.53	4.62	4.73	4.14	3.95	4.41
SG	3.10	3.55	5.23	5.21	4.66	4.71	4.31	3.84	4.13	4.10	3.92	4.16
SH	2.78	2.86	3.88	3.72	3.88	4.01	4.04	3.97	4.04	4.11	3.97	4.08
SI	2.79	2.72	3.83	3.77	3.97	3.82	3.94	4.26	3.97	4.02	3.94	3.96
SJ	3.02	3.07	4.51	4.69	4.49	4.36	4.21	4.22	3.91	3.33	3.99	4.71
SK	2.83	2.86	3.83	3.94	3.97	3.82	3.87	3.29	4.03	3.86	3.94	4.11
SL	2.79	2.72	3.81	3.91	3.97	3.93	4.15	4.38	4.54	4.17	3.92	4.13
SM	2.79	2.86	3.83	3.77	4.00	3.86	4.02	4.27	4.04	4.11	3.93	4.00
Avg.	2.85	2.90	4.06	4.07	4.12	4.06	4.08	4.06	4.15	3.97	3.95	4.18
s.d.	0.11	0.25	0.46	0.49	0.27	0.31	0.23	0.37	0.27	0.25	0.03	0.22
Impr.		0.05		0.02		–0.06		–0.02		–0.18		0.23
s.d.		0.15		0.12		0.10		0.30		0.25		0.21
t		1.03		0.39								3.44
p		n.s.		n.s.								

alveolar stops (/t, d/) more than the labial and velar stops (/p, b, k, g/), as in the case of CV and VC syllables.

The response times for CV and VC syllables are listed in Table 5.7 and a plot of the response times averaged across the subjects is given in Figure 5.10. The response time increased with decrease in SNR, but there was no significant effect of processing on response time.

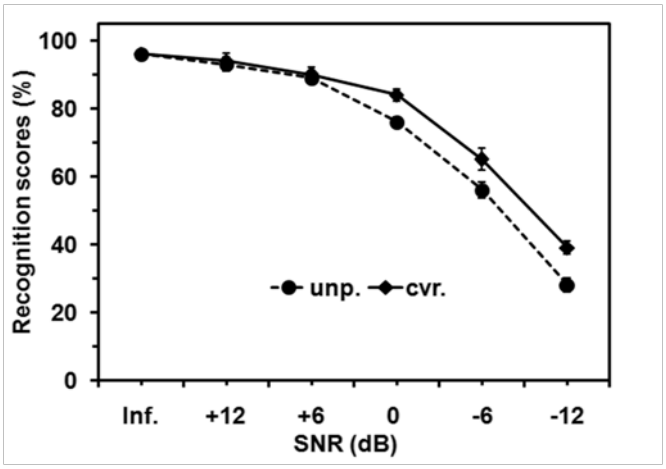


Figure 5.9 Experiment IV: Recognition scores (%) averaged across subjects for MRT wordlist. Error bars indicate standard deviations.

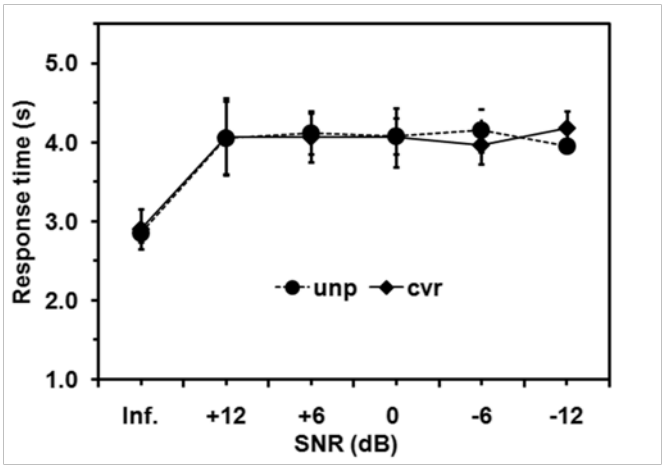


Figure 5.10 Experiment IV: Response times (s) averaged across subjects for MRT wordlist. Error bars indicate standard deviations.

5.4 Real-time implementation of CVR modification

To verify the feasibility of using the proposed technique in communication devices and hearing aids, it was implemented for real-time processing using a DSP board based on a 16-bit fixed-point processor with on-chip FFT hardware. The outputs of the real-time and offline processing were compared by examining the waveforms, listening, and an objective measure.

5.4.1 Implementation

The implementation was carried out on the DSP board "Spectrum Digital eZdsp" which is based on 16-bit fixed point processor "TI TMS320C5515" (Texas Instruments Inc., 2011; Spectrum Digital Inc., 2010; Gan et al., 2011). The board has 4 MB flash memory for user program and programmable stereo audio codec "TI TLV320AIC3204" (Texas Instruments Inc., 2008). The processor can operate up to a clock frequency of 120 MHz and has 16 MB address space with 320 KB on-chip RAM including 64 KB dual-access data RAM. Its other important features include DMA

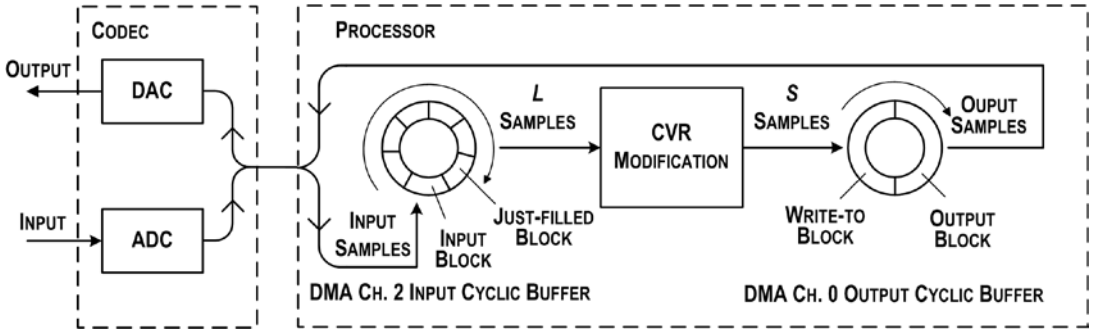


Fig. 5.11 Block diagram of implementation of automated CVR modification on DSP board.

controllers, 32-bit timers, and on-chip FFT hardware accelerator supporting up to 1024-point FFT computation. The program was written in C using "TI Code Composer Studio v 4.0". The processor clock frequency was set at 120 MHz and only one channel of the stereo codec was used with 16-bit quantization and sampling frequency of 10 kHz.

Figure 5.11 shows the block diagram of the implementation. The data transfer and buffering operations are interrupt driven and have been devised for an efficient realization of the processing with analysis frame length of 6 ms and frame shift of 1 ms. The input-output operations are handled using two DMA channels and two cyclic buffers. The input and output cyclic buffers have 7 and 2 blocks, respectively, with the block size *S* of 10 samples corresponding to 1 ms. At the set sampling frequency, DMA channel-2 reads the input samples from ADC into the current block of the input cyclic buffer and DMA channel-0 writes the processed samples from the current block of the output cyclic buffer to DAC. Cyclically incremented pointers keep track of current input, just-filled input, current output, and write-to output blocks. A 512-point buffer initialized with zero values is used as the data buffer. When the current input block gets filled, DMA interrupt is generated. At each interrupt, frame length *L* of 60 samples from six blocks of the input cyclic buffer are transferred to the data buffer, the processed *S* samples are transferred to the write-to block of the output cyclic buffer, and the pointers are updated.

The processing steps are the same as shown in Fig. 5.1, with due care of the constraints of fixed-point arithmetic, use of cyclic buffers for realizing delay lines, and utilization of the processor features to complete the processing of each frame within the frame shift duration. The energy *E* of the current frame is calculated and stored in a 20-sample cyclic buffer. The mean value of these samples is calculated as the smoothed energy E_s. The peak energy E_p is calculated using (5.3) with $\alpha = 255/256$ as an approximation to $(0.5)^{1/200}$, and stored in a 20-sample cyclic buffer. The peak energy E_p is used for calculating the scaling factor for the centroid-stabilizing tone to be added to the signal. The pre-stored samples of the tone with energy E_t are multiplied by $\beta = 0.1(E_p/E_t)^{0.5}$ and added to the signal samples. A Hanning window is applied on the frame, and the magnitude spectrum

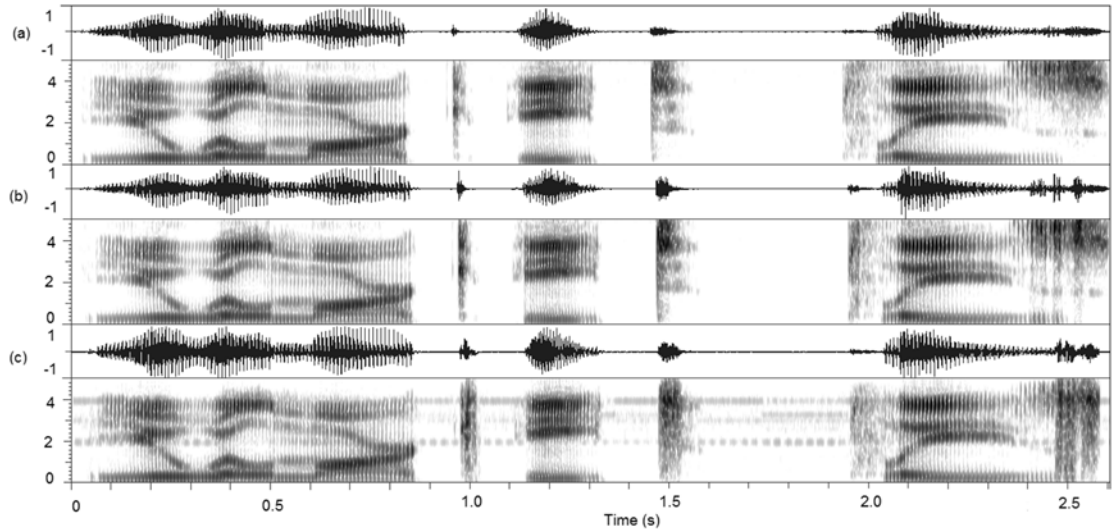


Figure 5.12 Example of offline and real-time processing for CVR modification: (a) speech signal of the utterance "*you will mark ut please*" and its spectrogram, (b) offline processed output and its spectrogram, (c) real-time processed output and its spectrogram. Frequency axis of spectrogram in kHz.

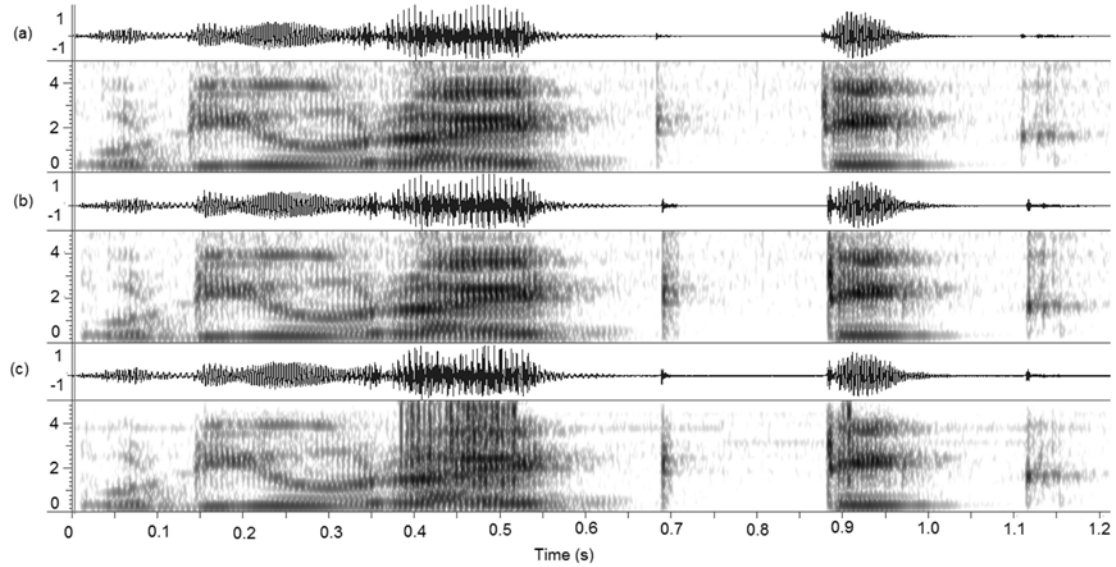


Figure 5.13 Example of offline and real-time processing for CVR modification: (a) speech signal of the utterance "*would you write tick*" and its spectrogram, (b) offline processed output and its spectrogram, (c) real-time processed output and its spectrogram. Frequency axis of spectrogram in kHz.

is calculated using 512-point FFT and stored in a 20-frame circular buffer. Smoothed spectrum is calculated by ensemble averaging and is used to calculate the centroid which is stored in a 20-sample circular buffer. A 20-point median of these values is calculated as the centroid F_c of the current frame, stored in a 20-sample circular buffer, and used to calculate 20-point first difference. The value of the

CVR modification flag is determined using hysteresis comparison as given in (5.2) and is used in calculating the gain factor G using (5.3), (5.4), and (5.5).

The last step of the processing involves multiplication of the ten samples of the input with the gain factor and outputting them. The delay in the signal path to compensate for the delay in the detection of spectral transitions is realized using a 10-block cyclic buffer. A scaling factor of 2^6 is used during gain calculation for improving the precision during fixed-point arithmetic. The same factor is used to scale down the values after multiplication of the delayed input samples with the gain factor. The processing involves algorithmic delay of 10 ms and computational delay of 1 ms.

5.4.2 Verification

The implementation for real-time processing was verified by observing the input and output waveforms on a digital storage oscilloscope. The signal delay (including algorithmic, computational, and input-output delays) was found to be nearly 21 ms. For a detailed comparison of offline and real-time processing, a PC was used to apply the speech stimuli used in the listening tests as input to the DSP board and to acquire its output. Two examples of the processing are shown Figures 5.12 and 5.13. There was no perceptible difference between the outputs of offline and real-time processing for all the test sentences.

Processing for CVR modification involves application of a time-varying gain function on the input signal, and therefore correlation between short-time energy envelopes can be used as an objective measure of the closeness of the outputs from two implementations. For this purpose, the short-time energy envelopes were calculated using 20 ms rectangular window and 1 ms window shift. This evaluation was carried out using the speech material of Experiment III consisting of 36 utterances with CV and VC keywords. The correlations of the short-time energy envelopes were calculated for three pairs of time-aligned waveforms: (i) unprocessed and offline processed (up-off), (ii) unprocessed and real-time processed (up-rt), and (iii) offline processed and real-time processed (off-rt). The mean values of the correlations for the up-off, up-rt, and off-rt pairs were 0.94, 0.94, and 0.98, respectively. These values indicate a close similarity between the outputs of offline and real-time processing. The correlations were also calculated for waveform segments corresponding to the keywords and the values were found to be the same as for the full utterances.

5.5 Discussion

A method for CVR modification with a gain of up to 9 dB during segments with sharp spectral transitions as detected on the basis of change in spectral centroid has been presented. Its effectiveness in improving speech perception under adverse listening conditions was evaluated by conducting listening tests for consonant recognition in nonsense CV and VC syllables and MRT wordlist. The tests were conducted on normal-hearing subjects with speech-spectrum shaped noise as a masker. For

CV and VC utterances, processing resulted in 7 − 21% improvement in recognition scores at SNRs below 0 dB. The corresponding improvements for MRT wordlist were 9 − 10%. The processing for CVR modification resulted in an SNR advantage of 6 dB, 5 dB, and 3 dB for CV syllables, VC syllables, and MRT wordlist, respectively. CVR modification was equally effective for word initial and word final consonants. The analysis of recognition scores at different SNRs for individual consonants indicated CVR modification to be more effective for alveolar stops (/d, t/) than for labial and velar stops (/p, b, g, k/), which may be attributed to better burst detection in case of alveolar stops. No significant change in the response time was observed for the CVR modified stimuli indicating that the processing improved intelligibility without affecting perceptual load.

From the results of the listening tests, it may be concluded that CVR modification using the proposed method is helpful in improving speech perception by improving consonant recognition in the presence of spectrally shaped noise as a masker. It has an algorithmic delay of approximately 10 ms and it has a low computational complexity. It was implemented for real-time processing using a DSP board based on a 16-bit fixed-point processor with on-chip FFT hardware. The outputs from the real-time processing closely matched to those from offline processing and the signal delay introduced by the processing is acceptable for its application in hearing aids and communication devices. Thus the investigations have shown that the proposed technique for CVR modification may be used for processing of the speech signal to improve its perception under adverse listening conditions.

Chapter 6

SUMMARY AND CONCLUSIONS

6.1 Introduction

The research objective was to devise a signal processing technique based on the properties of clear speech for improving perception of stop consonants under adverse listening conditions and suitable for use in speech communication devices and hearing aids. Investigations were performed on signal processing techniques for modification of specific speech segments with algorithmic and computational delays compatible with real-time processing. The method assumes clean speech to be available and processing is performed to make it robust towards further degradations under adverse listening conditions.

Investigations were performed on automated detection of landmarks associated with stop consonants in continuous speech. The effectiveness of parameters, distance measures, and time-steps on landmark detection were investigated. Methods suited for automated detection of stop consonant landmarks were identified and techniques for CVR modification and time-scale modification were developed for improving recognition of consonants. Intelligibility advantages of the two techniques were evaluated by conducting listening tests on normal-hearing subjects with speech-spectrum shaped noise as a masker. The method of CVR modification was implemented and tested for satisfactory real-time operation on a DSP board. The summary of the investigations, conclusions drawn on the basis of results and some suggestions for further research are given in the following sections.

6.2 Summary of the investigations

The research reported in this thesis can be summarized as the following.

1) *Landmark detection for speech intelligibility enhancement*

A landmark detector for use in a speech intelligibility enhancement application should detect the landmarks of interest with good temporal accuracy, preferably with low insertion rates, and with limited contextual information. The algorithmic and computational delays involved should be compatible with the constraints of real-time implementation. The

objective of the investigation on landmark detection was to derive an effective set of parameters, distance measure, and time-step, for detection of stop consonant landmarks in continuous speech. Four methods were investigated for the detection of the landmarks: (i) subband energies and centroids (EC), (ii) parameters of Gaussian mixture model (GMM) (iii) spectral moments (SM), and (iv) spectral moments with a tone added to the speech signal (SMTA). The voicing offset and the voicing onset landmarks were detected using energy variations in the band 0 − 400 Hz. The performance of the landmark detection methods was evaluated using manually annotated speech material involving 180 VCV utterances and 50 conversational style sentences from the TIMIT database.

Out of the four methods investigated, best detection rates were obtained for the GMM method, which involved excessive computations for parameter estimation. The method SM performed better than the method EC, but involved more computations during parameter estimation. All the three methods EC, GMM, and SM have algorithmic delay of nearly 400 ms and thus are not suited for real-time implementation which requires that the delays should not exceed 30 ms in order to maintain the synchronism between audio and visual channels. The method SMTA using rate of change of centroid derived from tone-added speech signal has detection rates lower than the other methods. However, it is much less computation intensive and involves algorithmic delay of approximately 10 ms. Therefore it can be considered as suited for real-time detection of burst and frication onset landmarks in continuous speech.

2) *Automated speech intelligibility enhancement using CVR and time-scale modification*

Investigations were made on automated enhancement of speech intelligibility by CVR and time-scale modification. The regions for modification were located using the landmark detection method based on spectral moments (SM). Experiments were conducted for CVR modification (Exp. I) and time-scale modification (Exp. II) using isolated VCV utterances involving 6 stop consonants (/b, d, g, p, t, k/) paired with 3 vowels (/a, i, u/) as the test material. Experiment I involved CVR modification by 9 dB. Experiment II involved expansion of the CV transition by a factor of 1.5 using sinusoidal model based analysis/synthesis. Both experiments were conducted on 5 normal-hearing subjects using a computerized test administration setup. The response choices and response times were recorded. The stimuli were mixed with speech-spectrum shaped noise at SNRs of 12, 6, 0, −6, −12 dB.

Results of CVR modification experiment (Exp. I) indicated statistically significant improvement in recognition scores by 7, 18, and 25% at SNRs of 0, −6, and −12 dB, respectively. The results of time-scale modification (Exp. II) indicated improvement in

recognition scores by 4, 7, and 7% at SNRs of 0, −6, and −12 dB, respectively. Improvements were not statistically significant. The improvements in recognition scores were equivalent to SNR advantages of 6 dB and 2 dB, for CVR modification and time-scale modification, respectively. There was no significant effect of either type of processing on the response times.

3) *Real-time speech intelligibility enhancement using CVR modification*

On the basis of results of Exp. I and Exp. II, it was concluded that CVR modification is well suited for real-time processing of speech signals for improving intelligibility under adverse listening conditions. Two sets of listening tests Exp. III and Exp. IV were conducted using an algorithm compatible with the requirements of real-time processing. The regions for modification were located using the landmark method based on spectral moments with tone-addition (SMTA). The test material for Experiment III involved 18 CV and 18 VC utterances recorded from a male speaker. There were 6 stop consonants (/b, d, g, p, t, k/) paired with 3 vowels (/a, i, u/) embedded in a carrier sentence "*you will mark* ----- *please*". Speech-spectrum shaped noise was used as masker at SNRs of 6, 3, 0, −3, −6, −9, and −12 dB. Experiment IV involved MRT with 300 CVC keywords with the carrier sentence "*would you write* -----" recorded from a male speaker. Speech-spectrum shaped noise was used as masker at SNRs of 12, 6, 0, −6, and −12 dB. Experiments III and IV were conducted on 10 normal-hearing subjects using a computerized test administration setup. The responses to the stimuli were tabulated in the form of stimulus-response matrices. The response times were also recorded.

For CV syllables, improvements in recognition scores of 8, 9, and 19% were obtained at 0, −6, and −12 dB SNRs. The corresponding improvements for VC syllables were by 9, 11, and 14%. The increase in recognition scores averaged across the subjects corresponded to an SNR advantage of 6 and 5 dB for the CV syllables and the VC syllables, respectively. The improvements were contributed mainly because of better reception of the place feature. Improvements in recognition scores of 8, 9, and 11% were obtained for MRT wordlist at SNRs of 0, −6, and −12 dB, respectively. The increase in recognition scores corresponded to an SNR advantage of 3 dB. The processing was equally effective for syllable initial and syllable final consonants.

The method for real-time CVR modification was implemented and tested for satisfactory operation on a DSP board with 16-bit fixed point processor "TI TMS320C5515" with on-chip FFT hardware. The technique has an algorithmic delay of 10 ms and its real-time implementation results in a signal delay of 21 ms.

6.3 Conclusions

From the results of the investigation, it may be concluded that CVR modification using the proposed method is helpful in improving speech perception by improving consonant recognition in the presence of spectrally shaped noise as a masker. It has an algorithmic delay of approximately 10 ms and it has a low computational complexity, making it suitable for real-time implementation. To verify the feasibility of using the proposed technique in applications involving real-time processing, it was implemented on a DSP board based on a 16-bit fixed-point processor with on-chip FFT hardware and the implementation was verified for satisfactory operation. The signal delay introduced by the processing was within the acceptable values for audio-visual delay. Therefore, the proposed technique for CVR modification can be used for processing of the speech signal to improve its perception under adverse listening conditions.

6.4 Suggestions for further research

The investigations on landmark detection in this thesis assume clear speech to be available for processing. This may limit the usefulness of speech enhancement techniques in practical situations where the input speech is noisy. In case of noisy input, the method may be used along with a speech enhancement technique for noise suppression (Loizou, 2007). As the output of noise suppression technique may have residual noise, study of the effect of additive noise on performance of landmark detection methods is of great practical significance and needs to be investigated. The possibility of improving performance of the EC method using auditory critical bands may be investigated. In case of the SMTA method, use of spectral centroids based on a mel-scale spectrum may also be useful for landmark detection, and it needs to be investigated. We have restricted our investigations to the detection of landmarks associated with stop consonants. Further investigations may be carried out for devising landmark detection techniques and intelligibility enhancement strategies for other classes of sounds that are critical to speech intelligibility. The investigations reported in the third chapter have shown that the GMM based landmark detection results in high temporal accuracy but the method is not suited for real-time applications because of its algorithmic delay and computational complexity. Its usefulness in speech recognition techniques involving landmark detection may be examined.

The investigations on automated enhancement of speech as reported in the fourth chapter have been concerned with improving speech perception under adverse listening conditions. The techniques involving CVR modification and duration modification used as pre-processing stages may be helpful in improving the performance of speech recognition systems. Although duration modification did not result in as high improvements in

recognition scores as CVR modification, the contributions of the two techniques appeared to be complementary. Therefore, it will be worthwhile to explore the use of these techniques for speech recognition.

The real-time CVR modification presented in the fifth chapter can be especially useful in wired or wireless two-way speech communication devices where the received speech signal is noise free and the noise gets acoustically added at the listener end. It can also be directly applicable for use in personal FM systems with hearing aids in the classrooms for hearing impaired students. It may also be useful for improving speech intelligibility of the output of public address systems with noisy listening environments. The effectiveness of the processing for these applications needs to be evaluated and implementation related issues need to be examined. Usefulness of the processing for improving speech intelligibility of the public announcement systems with significant reverberation and noise in listening environments may also be investigated. An interesting aspect of the announcement systems is that these systems need not use real-time processing and therefore an appropriate combination of CVR and duration modification can be used.

The real-time speech CVR modification technique presented in this thesis needs to be evaluated on subjects having sensorineural loss to establish the optimal value of CVR modification and to evaluate its effectiveness in improving speech intelligibility when combined with frequency-selective amplification and dynamic range compression (Dillon, 2001). Implementation of the technique on different types of processing platforms available in communication devices and hearing aids also needs to be investigated.

Appendix A

TEST INSTRUCTIONS AND FORMS

A.1 Introduction

Before the commencement of the test sessions, the subject was briefed on the objectives and usefulness of the listening tests in developing a speech processing scheme for improving speech perception under adverse listening conditions. The test instructions were given in writing and were also verbally explained. The tests were conducted after the subjects agreed to volunteer and signed the consent form.

A.2 Instructions for VCV test for CVR modification (Exp. I) and VCV test for time-scale modification (Exp. II)

1. You will be seated in front of a computer terminal with a mouse to click on the appropriate choice button on the computer screen. The sounds presented will be adjusted to your comfortable level. Be relaxed and attentive throughout the test. There is a trial test to become familiar with the procedure and sounds.
2. The display on the screen will show the following items
    - "*Play*" button to listen to the sound
    - "Response" panel with six choices to mark your choice
    - "*Next*" button to move to the next presentation
3. For every presentation, you will hear an isolated nonsense utterance in a vowel-consonant-vowel format. The vowel sound will be the same before and after the central consonant sound.
4. Your task is to click "*Play*" button to listen to the sound presented to you over the headphone and then click on the best matching sound among the six possible response choices displayed in the response panel. The sound will be presented once only and if you cannot recognize the sound, you have to guess. Click the "*Next*" button for listening to the next utterance. This procedure will be repeated until a set of 60 sounds are presented.
5. Each test will last about 10 minutes with a break of about 5 – 10 minutes between two consecutive tests. You may a take a maximum of six tests on a day. The total set of 180 tests

will be distributed over several days as per your convenience.. You can stop taking the tests at any time on a day and can discontinue from participating in tests before completion of the tests.

A.3 Instructions for CV/VC test (Exp. III)

1. You will be seated in front of a computer terminal with a mouse to click on the appropriate choice button on the computer screen. The sounds presented will be adjusted to your comfortable level. Be relaxed and attentive throughout the test. There is a trial test to become familiar with the procedure and sounds.
2. The display on the screen will show the following items
    - "*Play*" button to listen to the sound
    - "Response" panel with six choices to mark your choice
    - "*Next*" button to move to the next presentation
3. For every presentation, you will hear a vowel-consonant or consonant-vowel sound embedded in a carrier sentence "*You will mark ------ please*".
4. Your task is to click "*Play*" button to listen to the sound presented to you over the headphone and then click on the best matching sound among the six possible response choices displayed in the response panel. The sound will be presented once only and if you cannot recognize the sound, you have to guess. Click the "*Next*" button for listening to the next utterance. This procedure will be repeated until a set of 30 sounds are presented.
5. Each test will last about 10 minutes with a break of about 5 – 10 minutes between two consecutive tests. You may a take a maximum of six tests on a day. The total set of 96 tests will be distributed over several days as per your convenience. You can stop taking the tests at any time on a day and can discontinue from participating in tests before completion of the tests.

A.4 Instructions for MRT (Exp. IV)

1. You will be seated in front of a computer terminal with a mouse to click on the appropriate choice button on the computer screen. The sounds presented will be adjusted to your comfortable level. Be relaxed and attentive throughout the test. There is a trial test to become familiar with the procedure and sounds.
2. The display on the screen will show the following items
    - "*Play*" button to listen to the sound
    - "Response" panel with six choices to mark your choice
    - "*Next*" button to move to the next presentation
3. For every presentation, you will hear a vowel-consonant-vowel sound embedded in a carrier sentence "*Would you write ------*".

4. Your task is to click "*Play*" button to listen to the sound presented to you over the headphone and then click on the best matching sound among the six possible response choices displayed in the response panel. The sound will be presented once only and if you cannot recognize the sound, you have to guess. Click the "*Next*" button for listening to the next utterance. This procedure will be repeated until a set of 50 sounds are presented.
5. Each test will last about 15 – 20 minutes with a break of about 5 – 10 minutes between two consecutive tests. You may a take a maximum of three tests on a day. The total set of 72 tests will be distributed over several days as per your convenience. You can stop taking the tests at any time on a day and can discontinue from participating in tests before completion of the tests.

A. 5 Form for recording background information of the subjects

SUBJECT BACKGROUND INFORMATION

Date ___/___/____

Name _____ Code_____

Address _____

_____

Phone ( )_____ _____Extension _____

_____

Sex _____ Age_____

Occupation _____

Place of birth _____

First language _____

Other languages_____

History of noise exposure _____

History of hearing problems:_____

_____

Other remarks:_____

A. 6 Form for subject's willingness to participate

CONSENT FORM

I have carefully read the test instructions provided by A. R. Jayan (Ph.D. Scholar, IIT Bombay) for participation in listening experiments for evaluation of speech processing schemes. I am willing to participate in tests conducted by him. I understand that I can discontinue the participation at any time and that the data obtained from the tests will be used only for research without identifying me.

Signature: _____

Name: _____

Address: _____

_____

_____

Date ____/____/_____

Appendix B

TEST MATERIAL FOR MODIFIED RHYME TEST (MRT)

The test material used for the MRT (House, et al., 1965; ANSI, 1989) conducted for consonant recognition are listed in Table B.1 and B.2. The 300 CVC words used are arranged in 50 groups, each group consisting of six rhyming words. The groups listed in Table B.1 have different word-initial consonants and the groups in the Table B.2 have different word-final consonants.

Table B.1: Groups with different word-initial consonants

Group no.	Rhyming words					
01	went	sent	bent	dent	tent	rent
02	hold	cold	told	fold	sold	gold
03	kit	bit	fit	hit	wit	sit
04	must	bust	gust	rust	dust	just
05	bed	led	fed	red	wed	shed
06	pin	sin	tin	fin	din	win
07	not	tot	got	pot	hot	lot
08	vest	test	rest	best	west	nest
09	way	may	say	pay	day	gay
10	pig	big	dig	wig	rig	fig
11	shop	mop	cop	top	hop	pop
12	coil	oil	soil	toil	boil	foil
13	same	name	game	tame	came	fame
14	peel	reel	feel	eel	keel	heel
15	hark	dark	mark	bark	park	lark
16	thaw	law	raw	paw	jaw	saw
17	pen	hen	men	then	den	ten
18	heat	neat	feat	seat	meat	beat
19	dip	sip	hip	tip	lip	rip
20	hang	san	bang	rang	fang	gang
21	took	cook	look	hook	shook	book
22	fill	kill	will	hill	till	bill
23	bale	gale	sale	tale	pale	male
24	wick	sick	kick	lick	pick	tick
25	fun	sun	bun	gun	run	nun

Table B.2: MRT Groups with different word-final consonants

Group no.	Rhyming words					
01	pat	pad	pan	path	pack	pass
02	lane	lat	late	lake	lace	lame
03	teak	team	teal	teach	tear	tease
04	din	dill	dim	dig	dip	did
05	dug	dung	duck	dud	dub	dun
06	sum	sun	sung	sup	sub	sud
07	seep	seen	seethe	seek	seem	seed
08	pig	pill	pin	pip	pit	pick
09	back	bath	bad	bass	bat	ban
10	pale	pace	page	pane	pay	pave
11	cane	case	cape	cake	came	cave
12	tan	tang	tap	tack	tam	tab
13	fit	fib	fizz	fill	fig	fin
14	heave	hear	heat	heal	heap	heath
15	cup	cut	cud	cuff	cuss	cud
16	puff	puck	pub	pus	pup	pun
17	bean	beach	beat	beak	bead	beam
18	kill	kin	kit	kick	king	kid
19	mass	math	map	mat	man	mad
20	ray	raze	rate	rave	rake	race
21	save	same	sale	sane	sake	safe
22	sill	sick	sip	sing	sit	sin
23	peace	peas	peak	peach	peat	peal
24	bun	bus	but	bug	buck	buff
25	sag	sat	sass	sack	sad	sap

REFERENCES

ANSI. (**1989**). American National Standard Method for Measuring the Intelligibility of Speech over Communication Systems, ANSI/ASA S3.2-2009, American Standards Association, New York.

Apoux, F., Crouzet, O., and Lorenzi, C. (**2001**). " Temporal envelope expansion of speech in noise for normal-hearing and hearing-impaired listeners: effects on identification performance and response times," Hear. Res. **153**, 123–131.

Baer, T., Moore, B.C.J., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev. **30**, 49–72.

Boersma, P., and Weenink, D. (**1992**). Praat: doing phonetics by computer [Computer program]. Version 5.3.39, retrieved 2nd September 2005 from http://www.praat.org/.

Bradlow, A.R., and Pisoni, D.B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Bradlow, A.R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A.R., Kraus, N., and Hayes, E. (**2003**). "Speaking clearly for children with learning disabilities," J. Speech Lang. Hear. Res. **46**, 80–97.

Chen, F.R. (**1980**). Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level, M. S. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Colotte, V., and Laprie, Y. (**2000**). "Automatic enhancement of speech intelligibility," Proc. ICASSP 2000, Istanbul, Turkey, 1057–1060.

Deller, J.R., Hansen, J.H.L., and Proakis, J.G. (**2000**). Discrete-Time Processing of Speech Signals, (John Wiley, New York).

Dillon, H. (**2001**). Hearing Aids. New York: Thieme Medical.

Duda, R.O., Hart, P.E., and Stork, D.G. (**2004**). Pattern Classification, 2nd ed. (John Wiley, Singapore).

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

Gan, W.S., Seth, A., and Kuo, S.M. (**2011**). "Versatile and portable DSP platform for learning embedded signal processing," Proc. ICASSP 2011, Praugue, Czech Republic, 2888–2891.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., and Zue., V. (**1993**). TIMIT Acoustic-Phonetic Continuous Speech Corpus., Linguistic Data Consortium, Philadelphia.

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

Gordon-Salant, S. (**1987**). "Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects," J. Acoust. Soc. Am. **81**, 1199–1202.

Griffin, D.W., and Lim, J.S. (**1984**). "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoustics, Speech, Signal Process. **32**, 236–243.

Guelke, R.W. (**1987**). "Consonant burst enhancement: A possible means to improve intelligibility for the hard of hearing," J. Rehab. Res. Develop. **24**, 217–220.

Hazan, V., and Simpson, A. (**1998**). "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Commun. **24**, 211–226.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (**1965**). "Articulation testing methods: consonantal differentiation with closed-response set," J. Acoust. Soc. Am. **37,** 158–166.

Jayan, (**2014**), "Speech files used as the test material for evaluation of speech enhancement techniques", [online] Available: http://www.ee.iitb.ac.in/~spilab/material/jayan _phd2014"

Jayan, A.R., Pandey, P.C., and Lehana, P.K. (**2007**). "Time-scaling of consonant-vowel transitions using harmonic plus noise model for improving speech perception by listeners with moderate sensorineural impairment," Proc. 19th Int. Congress Acoustics (ICA 2007), Madrid, Spain, CAS-03-006.

Jayan, A.R., Pandey, P.C., and Lehana, P.K. (**2008**). "Automated detection of transition segments for intensity and time-scale modification for speech intelligibility enhancement," Proc. 19th IEEE Int. Conference on Signal Processing, Communications, and Networking (ICSCN 2008), Chennai, India, 63–68.

Jayan, A.R., and Pandey, P.C. (**2008**). "Automated detection of speech landmarks using Gaussian mixture modeling," Proc. Int. Symposium on Frontiers of Research on Speech and Music (FRSM 2008), Kolkata, India, 323−327.

Jayan, A.R., and Pandey, P.C. (**2009**). "Detection of stop landmarks using Gaussian mixture modeling of speech spectrum," Proc. ICASSP 2009, Taipei, Taiwan, 4681–4684.

Jayan, A.R., Rajath Bhat, P.S., and Pandey, P.C. (**2011**). "Detection of burst onset landmarks in speech using rate of change of spectral moments", Proc. 17th Nat. Conf. Communications (NCC 2011), Bangalore, India, Sp. Pr. I, P3.

Jayan, A.R., and Pandey, P.C. (**2012**). "Automated CVR modification for improving perception of stop consonants", Proc. 18th Nat. Conf. Communications (NCC 2012), Kharagpur, India, 698−702.

Kapoor, A., and Allen, J.B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**, 478−491.

Kates, J.M. (**1994**). "Speech enhancement based on a sinusoidal model," J. Speech. Hear. Res. **37**, 449−464.

Kennedy, E., Levitt, H., Neuman, A.C., and Weiss, M. (**1997**). "Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," J. Acoust. Soc. Am. **103**, 1098–1114.

Koning, R., and Wouters, J. (**2012**). "The potential of onset enhancement for increased speech intelligibility in auditory prostheses," J. Acoust. Soc. Am. **132**, 2569−2581.

Krause, J.C., and Braida, L.D. (**2002**). "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," J. Acoust. Soc. Am. **112**, 2165–2172.

Krause, J.C., and Braida, L.D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362−378.

Laroche, J., Stylianou, Y., and Moulines, E. (**1993**). "HNS: Speech modification based on a harmonic + noise model," Proc. ICASSP 1993, Minneapolis, Minnesota, 550–553.

Li, F., Menon, A., and Allen, J.B. (**2010**). "A psychoacoustic method to find the perceptual cues to stop consonants in natural speech," J. Acoust. Soc. Am. **127**, 2599−2610.

Li, F., Trevino, A., Menon, A., and Allen, J.B. (**2012**). "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," J. Acoust. Soc. Am. **132**, 2663−2675.

Li, N., and Loizou, P.C., (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 3947−3958.

Lin, C.Y., and Wang, H.C. (**2008**). "Mandarin stops classification using random forest approach," Proc. 6th International Symposium on Chinese Spoken Language Processing (ISCSLP 2008), Kunming, China, 241−245.

Liu, S., and Zeng, F.G. (**2006**). "Temporal properties in clear speech perception," J. Acoust. Soc. Am. **120**, 424–432.

Liu, S.A. (**1995**). Landmark detection for distinctive feature-based speech recognition. Ph.D. Thesis, MIT, Cambridge, MA.

Liu, S.A. (**1996**). “Landmark detection for distinctive feature-based speech recognition,” J. Acoust. Soc. Am. **100**, 3417–3430.

Loizou, P.C. (**2007**). Speech Enhancement: Theory and Practice. New York: CRC, 2007.

Mahalanobis, P.C. (**1936**). “On the generalized distance in statistics,” Proc. National Institute of Sciences of India. **2**, 49–55.

Malah, D., (**1979**). “Time-Domain algorithms for harmonic bandwidth reduction and time scaling of speech signals,” IEEE Trans. Acoustics, Speech, Signal Process. **27**, 121–133.

McAulay, R.J., and Quatieri, T.F. (**1986**). “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Trans. Acoustics, Speech, Signal Process. **34**, 744–754.

Miller, G.A., and Nicely, P.E. (**1955**). “An analysis of perceptual confusions among some English consonants,” J. Acoust. Soc. Am. **27**, 338–352.

Montgomery, A.A., and Edge, R.A. (**1988**). “Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults,” J. Speech Hear. Res. **31**, 386–393.

Moulines, E., and Laroche, J. (**1995**). “Non-parametric techniques for pitch-scale and time-scale modification of speech,” Speech Commun. **16**, 175–205.

Nejime, Y., Aritsuka, T, Imamura, T., Ifukube, T., and Matsushima, J. (**1996**). “A portable digital speech-rate converter for hearing impairment,” IEEE Trans. Rehabil. Engineering. **4**, 73−83.

Nejime, Y., and Moore, B.C.J. (**1998**). “Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss,” J. Acoust. Soc. Am. **103**, 572–576.

Nguyen, B.P., and Akagi, M. (**2009**). “A flexible spectral modification method on temporal decomposition and Gaussian mixture model,” J. Acoust. Sci. & Tech. **30**, 170−179.

Niyogi, P., and Sondhi, M.M. (**2002**). “Detecting stop consonants in continuous speech,” J. Acoust. Soc. Am. **111**, 1063–1076.

Ortega, M., Hazan, V., and Huckvale, M. (**2000**). “Automatic cue enhancement of natural speech for improved intelligibility,” Speech, Hearing, and Language: work in progress. **12**, 42–56.

Paliwal, K.K. (**1998**). “Spectral subband centroid features for speech recognition,” Proc. ICASSP 1998, Seattle, Washington, 617−620.

Pantazis, Y., and Stylianou, Y. (**2008**). “Improving the modeling of noise part in the harmonic plus noise model of speech,” Proc. ICASSP 2008, Las Vegas, Nevada, 4609−4612.

Park, C. (**2008**). Consonant landmark detection for speech recognition, Ph.D Thesis, MIT, Cambridge, MA.

Payton, K.L., Uchanski, R.M., and Braida, L.D. (**1994**). “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,” J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (**1985**). “Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech,” J. Speech Hear. Res. **28**, 96–103.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (**1986**). “Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech,” J. Speech Hear. Res. **29**, 434–446.

Picheny, M.A., Durlach, N.I., and Braida, L.D. (**1989**). “Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech,” J. Speech Hear. Res. **32**, 600–603.

Quatieri, T.F., and McAulay, R.J. (**1986**). “Speech transformation based on a sinusoidal representation,” IEEE Trans. Acoustics, Speech, Signal Process. **34**, 1449–1464.

Quatieri, T.F., and McAulay, R.J. (**1992**). “Shape invariant time-scale and pitch modification of speech,” IEEE Trans. Signal Process. **40**, 497–510.

Rasetshwane, D.M. (**2009**). Enhancement of speech intelligibility using speech transients extracted by a wavelet packet-based real-time algorithm. Ph.D. Thesis, University of Pittsburgh.

Regnier, M.S., and Allen, J.B. (**2008**). “A method to find noise-robust perceptual features: Application for consonant /t/,” J. Acoust. Soc. Am. **123**, 2801 – 2814.

Salomon, A., Espy-Wilson, C.Y., and Deshmukh, O. (**2004**). “Detection of speech landmarks: Use of temporal information,” J. Acoust. Soc. Am. **115**, 1296 – 1305.

Skowronski, M.D., and Harris, J.G. (**2006**). “Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments,” Speech Commun. **48**, 549 – 558.

Spectrum Digital, Inc. (**2010**). TMS320C5515 eZdsp USB Stick Technical Reference,” [online] available:support.spectrumdigital.com/boards/usbstk5515/reva/files/usbstk5515_TechRef_RevA.pdf.

Stevens, K.N. (**1981**). “Evidence for the role of acoustic boundaries in the perception of speech sounds”, J. Acoust. Soc. Am. **69**, S116.

Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S., and Liu, S. (**1992**). “Implementation of a model for lexical access based on features,” Proc. ICSLP 1992, Banff, Alberta, 499–502.

Stone, M.A., and Moore, B.C.J. (**2005**). "Tolerable hearing-aid delays: IV. effects on subjective disturbance during speech production by hearing-impaired subjects," J. Ear. Hear. Res. **26**, 225 – 235.

Stuttle, M.N., and Gales, M.J.F. (**2002**). "Combining a Gaussian mixture model front end with MFCC parameters," Proc. ICSLP 2002, Denver, Colorado, 1565 – 1568.

Stuttle, M.N. (**2003**). A Gaussian mixture model spectral representation for speech recognition. Ph.D. Thesis, University of Cambridge, Cambridge, UK.

Stylianou, Y. (**2001**). "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. Acoustics, Speech, Audio Process. **9**, 21 – 29.

Stylianou, Y. (**2005**). "Modeling speech based on harmonic plus noise models," in G. Chollet et al., (Eds.), in Nonlinear Speech Modeling, Springer-Verlag, Berlin. 244–260.

Tantibundhit, C., Pernkopf, F., and Kubin, G., (**2009**). "Speech enhancement based on joint time-frequency segmentation," Proc. ICASSP 2009, Taipei, Taiwan, 4673–4676.

Texas Instruments, Inc. (**2008**). "TLV320AIC3204 Ultra Low Power Stereo Audio Codec," [online] Available: focus.ti.com/lit/ds/symlink/tlv320aic3204.pdf.

Texas Instruments, Inc. (**2011**). "TMS320C5515 Fixed-Point Digital Signal Processor," [online] Available: http://focus.ti.com/lit/ds/symlink/tms320c5515.pdf.

Thomas, T.G. (**1996**). Experimental evaluation of improvement in speech perception with consonantal intensity and duration modification, Ph.D. Thesis, Dept. of Elect. Engg., IIT Bombay, Mumbai, India.

Uchanski, R.M., Choi, S.S., Braida, L.D., Reed, C.M., and Durlach, N.I. (**1996**). "Speaking clearly for the hard of hearing IV: Further studies on the role of speaking rate," J. Speech Hear. Res. **39**, 494–509.

Vaughan, N.E., Furukawa, I., Balasingam, N., Mortz, M., and Fausti, S.A. (**2002**). "Time-expanded speech recognition in older adults," J. Rehab. Res. Dev. **39**, 559–566.

Yoo, S.D., Boston, J.R., El-Jaroudi, A., and Li, C.C. (**2007**). "Speech signal modification to increase intelligibility in noisy environments," J. Acoust. Soc. Am. **122**, 1138–1149.

Zolfaghari, P., and Robinson, T. (**1996**). "Formant analysis using mixtures of Gaussians," Proc. ICSLP 1996, Philadelphia, PA, 1229–1232.

Zolfaghari, P., Kato, H., Minami, Y., Nakamura, A., Katagiri, S., and Patterson, R. (**2006**). "Dynamic assignment of Gaussian components in modelling speech spectra," J. VLSI Signal Proc. **45**, 7–19.

AUTHOR'S RESUME

A. R. Jayan received the B.Tech. degree in Electronics and Communication Engineering from University of Calicut in 1992. He received the M.Tech. degree in Digital Electronics from Cochin University of Science and Technology in 1994. He joined as a research engineer at Computer and Communications division of Electronics Research & Development Centre of India (ER&DCI) Thiruvananthapuram in 1994. In 1999, he joined the Department of Technical Education, Govt. of Kerala and is currently working as Associate Professor in the Department of Electronics and Communication at Govt. Engineering College, Thrissur. He is currently pursuing Ph.D. in the Department of Electrical Engineering, Indian Institute of Technology Bombay, India. His research interests include digital signal processing, design of digital systems, and electronics instrumentation.

LIST OF PUBLICATIONS

Journal Paper (**under review/revision**)

1. Jayan, A.R., and Pandey, P. C. (**2013**). "Automated consonant-vowel ratio modification for improving speech perception,"

Papers in International Conference Proceedings

1. Jayan, A.R., Pandey, P.C., and Lehana, P.K. (**2007**). "Time-scaling of consonant-vowel transitions using harmonic-plus-noise model for improving speech perception by listeners with moderate sensorineural impairment," Proc. 19th Int. Congress Acoustics 2007, Madrid, Spain, paper. no. CAS-03-006.
2. Jayan, A.R., and Pandey, P.C. (**2009**). "Detection of stop landmarks using Gaussian mixture modeling of speech spectrum," Proc. ICASSP 2009, Taipei, Taiwan, pp. 4681−4684.

Papers in National Conference Proceedings

1. Jayan, A.R., Pandey, P.C., and Lehana, P.K. (**2008**). "Automated detection of transition segments for intensity and time-Scale modification for speech intelligibility enhancement," Proc. IEEE Int. Conf. Signal Processing, Communications, Networking 2008, Chennai, India, pp. 69−74.
2. Jayan, A.R., Pandey, P.C., and Pandey, V.K. (**2008**). "Detection of acoustic landmarks with high resolution for speech processing," Proc. 14th National Conf. Communications 2008, Mumbai, India, pp. 427−431.
3. Jayan, A.R., and Pandey, P.C. (**2008**). "Automated detection of speech landmarks using Gaussian mixture modeling," Proc. Int. Symposium on Frontiers of Research on Speech and Music 2008, Kolkata, India, pp. 323−327.

4. Jayan, A.R., Rajath Bhat, P.S., and Pandey, P.C. (**2011**). “Detection of burst onset landmarks in speech using rate of change of spectral moments,” Proc. National Conference on Communications 2011, Bangalore, India, paper SpPrI.3.

5. Jayan, A.R., and Pandey, P.C. (**2012**). “Automated CVR modification for improving perception of stop consonants,” Proc. National Conference on Communications 2012, Kharagpur India, pp. 698–702, paper no. 1569512651.

Patent

1. Pandey, P.C., Jayan, A.R., and Tiwari, N. (**2014**). “Method and system for consonant-vowel ratio modification for improving speech perception,” Indian Patent Application No. 739/MUM/2014, April 25, 2014.

ACKNOWLEDGMENTS

I take this opportunity to express my deep sense of gratitude and profound respect to my supervisor Prof. P. C. Pandey for the unconditional support, invaluable guidance, motivation, and constructive criticism which have made this work possible. I was always inspired by his wisdom, insightful nature, curiosity, and strive for perfection and my gratitude towards him is beyond words. I am deeply thankful to Prof. Preeti Rao and Prof. V. M. Gadre, members of the research progress committee for their valuable suggestions and encouragements at various stages of my research work.

I would like to thank all members in the Signal Processing and Instrumentation Laboratory (SPI lab), EE Dept., IIT Bombay for their support and encouragement. I am grateful to Alice, Vinod, Parveen, Milind, Pandurang, Mohan, Nitya, and Vidyadhar for providing me support whenever I needed. I am thankful to several M.Tech. students in the lab, especially to Rajath, Jagbandhu, Nataraj, and Santosh for the company and discussions.

I remember my mother who taught me to read and write, and who left me halfway through this journey. I am deeply indebted to my father for planting the seeds of research in my mind. I am thankful to my in-laws, relatives, especially to Mr. Santhosh Kumar, for extending unconditional support during my stay at IIT Bombay. Words cannot express my gratitude towards my wife Malini, for refilling me with energy whenever I felt exhausted. I am thankful to my children Krishnanand and Jaswanth who always got puzzled by the nonsense sounds from my computer. Ultimately, I am thankful to God for always being with me during this journey.

A. R. Jayan

"Wisdom is not a product of schooling but of the lifelong attempt to acquire it"

–Albert Einstein