Speech and noise analysis using sparse representation and acoustic-phonetics knowledge

Thesis · December 2017

2 authors:

A.G. Ramakrishnan
Indian Institute of Science
369 PUBLICATIONS **2,768** CITATIONS
SEE PROFILE

K V Vijay Girish
Indian Institute of Science
14 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Speech and noise analysis using sparse representation and acoustic-phonetics knowledge

A Thesis
Submitted for the Degree of
Doctor of Philosophy
in the **Faculty of Engineering**

by

K Venkata Vijay Girish

Department of Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (INDIA)

April 2017

DEDICATED TO

My parents

Signature of the Author: .....

K Venkata Vijay Girish
Dept. of Electrical Engineering
Indian Institute of Science, Bangalore

Signature of the Thesis Supervisor: .....

Prof. A G Ramakrishnan
Professor
Dept. of Electrical Engineering
Indian Institute of Science, Bangalore

Acknowledgements

I am finally ready with my thesis after 6.5 years and few months. It has been a long journey with varied experiences. I have always wanted to join IISc which was my dream institute for higher studies, although initially I was not sure about which field to pursue. I am now very glad that I chose to work in a field which is a good blend of signal processing and machine learning. I want to thank my advisor Prof. A G Ramakrishnan who on the first day of meeting itself encouraged me to upgrade my registration to direct Ph.D. He really supported me during my whole stay at IISc, providing me guidance and ideas to go deeper into the problem, and teaching me the nuances of how to do research. I sincerely want to thank Dr. T V Ananthapadmanabha, an eminent researcher and entrepreneur for advising me on various directions to pursue speech research. I thank him for spending hours with me guiding me thoroughly and teaching me the intricacies of speech research. I thank the faculty of Electrical, Electrical Communication Engineering and Department of Computer Science and Automation for the excellent courses they offered. I thank MILE Lab for the excellent computational and other facilities. I thank the staff of the Electrical Engineering for their support and prompt help when required. I want to thank the funding agencies MHRD, IISc, DRDO and DST for funding travel to various workshops and conferences abroad, which gave me an opportunity to explore very nice destinations like Florence and San Francisco.

I thank my labmates Suresh, Jobin, Abhiram, Prathosh, Deepak, Anoop, Maj. Dabhi, Maj. Saurabh, Madhav and Pandey who provided a very encouraging and intellectual environment in the lab. I have made very good friends in IISc campus, whom I will forever remember. I cherish the nice discussions and long walks in and around the campus with Sunder Ram, who is a very easy going and a cheerful friend. I will remember the cheerful Karthik, my badminton partner; Neeraj who is very calm and a true researcher for discussions on various topics; Harikiran, Sai, Jitendra and Shiva for always supporting and encouraging me during my good and bad times. I want to thank Sai and Neeraj for accompanying me for the Goa trip and having a memorable time. A-mess has been the meeting place for various gossips and discussions; I will really miss the A-mess. I am grateful to the very good facilities in Gymkahana I have used and the various clubs: Aerobics, athletics and fitness, Kung-fu, swimming and music club for helping me maintain my fitness and learn new skills. I want to thank Mani and Nanda, my swimming coaches for their awesome training sessions in swimming pool, ground and gym. I want to thank the IISc administration for organizing various talks and events like

IISConnect, providing me a platform to interact with various industries and getting a job.

I want to thank my parents and sister who have always been supportive of me during my stay at IISc.

Abstract

This thesis addresses different aspects of machine listening using two different approaches, namely (1) A supervised and adaptive sparse representation based approach for identifying the type of background noise and the speaker and separating the speech and background noise, and (2) An unsupervised acoustic-phonetics knowledge based approach for detecting transitions between broad phonetic classes in a speech signal and significant excitation instants called as glottal closure instants (GCIs) in voiced speech, for applications like speech segmentation, recognition and modification.

Real life speech signals generally contain a foreground speech by a particular speaker in the presence of a background environment like factory or traffic noise. These audio signals termed as noisy speech signals are available in the form of recordings say, audio intercepts or real time signals which can be single channel or multi channel. Real time signals are available during mobile communication and in hearing aids. Processing of these signals has been approached by the research community for various independent applications like classification of components of the noisy speech signal, source separation, enhancement, speech recognition, audio coding, duration modification and speaker normalization. Machine listening encapsulates solutions to these applications in a single system. It extracts useful information from noisy speech signals, and attempts to understand the content as much as humans do. In the case of speech enhancement, especially for the hearing impaired, the suppression of background noise for improving the intelligibility of speech would be more effective, if the type of background noise can be classified first. Other interesting applications of noise identification are forensics, machinery noise diagnostics, robotic navigation systems and acoustic signature classification of aircrafts or vehicles. Another motivation to identify the nature of background noise is to narrow down to the possible geographical location of a speaker. Speaker classification helps us to identify the speaker in an audio intercept.

In the supervised sparse representation based approach, a dictionary learning based noise and speaker classification algorithm is proposed using a cosine similarity measure for learning atoms of the dictionary and is compared with other non-negative dictionary learning methods. For training, we learn dictionaries for speaker and noise sources separately using the various dictionary learning methods. We have used the Active Set Newton Algorithm (ASNA) and supervised non-negative matrix factorization for source recovery in the testing phase. Based on the objective measure of signal to distortion ratio (SDR), we get the frame-wise noise classification accuracy of 97.8% for fifteen

different noises taken from the NOISEX database. The proposed evaluation metric of sum of weights (SW) applied on concatenated dictionaries gives a good accuracy, for speaker classification on clean speech, using high energy subsets of test frames and dictionary atoms. We get the best utterance level speaker classification accuracy of 100% for 30 speakers taken from TIMIT database on clean speech.

We have then dealt with noisy speech signals assuming a single speaker speaking in a noisy environment. The noisy speech signals have been simulated at different SNRs using different noise and speaker sources. We have classified the speaker and background noise class of the noisy speech signal and subsequently separated the speech and noise components. Given a test noisy speech signal, a noise label is assigned to a subset of frames selected using the SDR measure, and an accumulated measure is used to classify the noise in the whole test signal. The speaker is classified using the proposed metric of accumulated sum of weights on high energy features, estimated using ASNA with L_1 regularization from the concatenation of speaker dictionaries and the identified noise source dictionary. Using the dictionaries of the identified speaker and noise source, we obtain the estimate of the separated speech and noise signal using ASNA with L_1 regularization and supervised non-negative matrix factorization (NMF). We obtain around 98% accuracy for noise classification and 89% for speaker classification at an SNR of 10 dB for a combination of 30 speakers and 15 noise sources.

In the case of an unknown noise, the noise source is estimated as the nearest known noise label. The distribution of an unknown noise source amongst the known noise classes gives an indication of the possible noise source. The dictionary corresponding to the estimated noise label is updated adaptively using the features from the noise-only frames of the test signal. The updated dictionary is then used for speaker classification, and subsequently separation is carried out. In the case of an unknown speaker, the nearest speaker is estimated and the corresponding dictionary is updated using a clean speech segment from the test signal. We assume that a clean speech segment is available for adapting the speech dictionary. We have observed an improvement in signal to distortion ratio (SDR) after separation of speech and noise components using an adaptive dictionary. Adaptive noise dictionary gives an improvement of about 18% in speaker classification accuracy and 4 dB in SDR over an out-of-set dictionary, after enhancement of noisy speech at an SNR of 0 dB.

In the case of a conversation, a divide and conquer algorithm is proposed to recursively estimate the noise sources, and estimate the approximate instant of noise transition and the number of noise types. We have then experimented on a conversation simulated by concatenating two different noise signals, each containing speech segments of distinct speakers and obtained a mean absolute error in the detection of noise transition instant of 10 ms at -10 dB SNR. Each of the segments obtained based on the transition instant can be treated as a single noise mixed with speech from a single speaker and subsequent speaker classification and source separation can be done as in the previous case.

We have also addressed the classification of speakers and subsequent separation of speakers in overlapped speech, obtaining a mean speaker classification accuracy of 84% for the speaker 1 to speaker 2 ratio (S1S2R) of 0 dB.

The advantage of the proposed dictionary learning and sparse representation based approach is

that the training and classification model is independent of the selected classes of speakers and noises. Dictionaries for new classes can be easily added or the old classes can be removed or replaced instead of retraining. Also, the same model can be used for identifying other types of classes like language and gender. We have achieved speaker and noise classification and subsequent separation using only spectral features for dictionary learning. This is in contrast to the stochastic model based approaches where the model needs to be retrained whenever a new class is added.

In the unsupervised acoustic-phonetics knowledge based approach, we detect transitions between broad phonetic classes in a speech signal which has applications such as landmark detection and segmentation. The proposed rule based hierarchical method detects transitions from silence to non-silence, sonorant to non-sonorant and vice-versa. We exploit the relative abrupt changes in the characteristics of the speech signal to detect the transitions. Relative thresholds learnt from a small development set are used to determine the parameter values. We propose different measures for detecting transitions between broad phonetic classes in a speech signal based on abrupt amplitude changes. A measure is defined on the quantized speech signal to detect transitions between very low amplitude or silence (S) and non-silence (N) segments. The S-segments could be stop closures, pauses or silence regions at the beginning and/or ending of an utterance. We propose two other measures to detect the transitions between sonorant and non-sonorant segments and vice-versa. We make use of the fact that most sonorants have higher energy in the low frequencies, than other phone classes such as unvoiced fricatives, affricates and unvoiced stops. For this reason, we use a bandpass speech signal (60-340 Hz) for extracting temporal features. A subset of the extrema (minimum or maximum amplitude samples) between every pair of successive zero-crossings and above a threshold is selected from each frame of the bandpass filtered speech signal. Occurrences of the first and the last extrema lie far before and after the mid-point (reference) of a frame, if the speech signal belongs to a non-transition segment; else, one of these locations lie within a few samples from the reference, indicating a transition frame. The advantage of this approach is that it does not require significant training data for determining the parameters of the proposed approach.

When tested on the entire TIMIT database for clean speech, of the transitions detected, 93.6% are within a tolerance of 20 ms from the hand labeled boundaries. Sonorant, unvoiced non-sonorant and silence classes and their respective onsets are detected with an accuracy of about 83.5% for the same tolerance using the labelled TIMIT database as reference. The results are as good as, and in some respects better than the state-of-the-art methods for similar tasks. The proposed method is also tested on the test set of the TIMIT database for robustness with respect to white, babble and Schroeder noise, and about 90% of the transitions are detected within the tolerance of 20 ms at the signal to noise ratio of 5 dB.

We have also estimated glottal closure instants (GCIs) useful for a variety of applications such as pitch and duration modification, speaking rate modification, pitch normalization, speech coding/ compression, and speaker normalization. The instant at which the vocal tract is significantly excited within each glottal cycle in a speech signal is referred to as the epoch or the GCI. Subband analysis

of linear prediction residual (LPR) is proposed to estimate the GCIs from voiced speech segments. A composite signal is derived as the sum of the envelopes of the subband components of the LPR signal. Appropriately chosen peaks of the composite signal are the GCI candidates. The temporal locations of the candidates are refined using the LPR to obtain the GCIs, which are validated against the GCIs obtained from the electroglottograph signal, recorded simultaneously. The robustness is studied using additive white, pink, blue, babble, vehicle, HF channel noises for different signal to noise ratios and reverberation. The proposed method is evaluated using six different databases and compared with three state-of-the-art LPR based methods. The GCI detection performance of the proposed algorithm is quantified using the following measures: identification rate (IDR), miss rate (MR), false alarm rate (FAR), standard deviation of error (SDE) and accuracy to 0.25 ms. We have shown that significant GCI information exists in each subband of speech up to 2000 Hz, and a minimum of 89% identification rate (for subbands other than lowpass) can be obtained for clean speech using the proposed method. The results show that the performance of the proposed method is comparable to the best of the LPR based techniques for clean, and noisy speech.

Abbreviations

Acc.25	accuracy to 0.25 ms
AGR	proposed algorithm for the detection of transitions
AM-FM	amplitude modulation- frequency modulation
ASNA	active-set Newton algorithm
ASNA-L1	active-set Newton algorithm with L_1 regularization term
between-CS	between-class cosine similarity
BPF	bandpass filtered
CMU	Carnegie Mellon University
CS	composite signal
D	order of each filter in the Hamming filterbank
dEGG	derivative of the electroglottograph signal
DESA	Discrete-time Energy Separation Algorithm
DF	distinctive feature
DL	dictionary learning
DYPSA	Dynamic Programming Phase Slope Algorithm
EER	equal error rate
EGG	electroglottograph signal
ERB	equivalent rectangular bandwidth
FAR	false alarm rate
GCI	glottal closure instant
GT	Gammatone
GTE	Gammatone subband envelope
H	high
HAM	Hamming subband AM-FM decomposition method
HBE	Hamming bandpass envelope
HBEVAR	HBE with variable center frequency
HL	high-low
HMM	hidden Markov model
IDR	identification rate
ILPR	integrated linear prediction residual
kHz	kilohertz
KL	Kullback-Leibler
L	low
LH	low-high
LP	linear prediction
MADE	mean absolute difference between extrema in a frame
MAE-S1S2R	mean absolute error between the original and estimated S1S2R
MAE-SNR	mean absolute error between the original and estimated SNR
mag.STFT	magnitude of the short-time Fourier transform
MFCC	mel frequency cepstral coefficient
MR	miss rate

N	non-silence
NDR	noise to distortion ratio
No.	number
NMF	non-negative matrix factorization
PCHIP	piecewise cubic hermite interpolating polynomial
OFE	occurrence of the first extremum
OLE	occurrence of the last extremum
PF	phonetic feature
PLC	percentage of larynx cycles
PSD	power spectral density
RIR	room impulse response
S	silence
S1S2R	ratio of energy in decibel of speech from speaker 1 (S1) to speaker 2 (S2) in overlapped speech
S1DR	ratio of energy in decibel of speech to the distortion between original and estimated speech from speaker 1 (S1)
S2DR	ratio of energy in decibel of speech to the distortion between original and estimated speech from speaker 2 (S2)
SDE	standard deviation of the timing error
SDR	signal to distortion ratio
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SI	silence index
SNMF	sparse NMF
SNR	signal to noise ratio
STD-S1S2R	standard deviation of error in the estimate of S1S2R
STD-SNR	standard deviation of error in the estimate of SNR
sup.NMF	source recovery using supervised NMF
SVM	support vector machine
T	threshold on ADE
T60	time required for magnitude of the RIR to decay by 60 dB
TDCS	Threshold dependent cosine similarity based dictionary learning algorithm
TEO	Teager Energy Operator
within-CS	within-class cosine similarity
YAGA	Yet Another GCI Algorithm
ZFF	zero frequency filtering

Notations

$\mathbf{A}$	active set
$a_k[n]$	AM component for the k^{th} subband signal
$a_r[i]$	reference amplitude
$C_p[n]$	envelope of the p^{th} subband signal
$C[n]$	composite signal
$cs_b(\mathbf{d}_n, \mathbf{d}_j)$	between-class cosine similarity between $\mathbf{d}_n \in \mathbf{D}^k$, $\mathbf{d}_j \in \mathbf{D}^h, k \neq h$
$cs_w(\mathbf{d}_n, \mathbf{d}_j)$	within-class cosine similarity between $\mathbf{d}_n, \mathbf{d}_j \in \mathbf{D}^k, n \neq j$
$\mathbf{D}$	generalized or concatenated dictionary
$\mathbf{D}^k$	dictionary for the k^{th} source
$\mathbf{D}^i_{sp}$	speech dictionary for the i^{th} source
$\mathbf{D}^j_{ns}$	noise dictionary for the j^{th} source
$\mathbf{d}^k_n$	n^{th} dictionary atom corresponding to $\mathbf{D}^k$
$\mathbf{D}_{upd}$	updated dictionary
$diag()$	diagonal matrix whose diagonal entries consists of its argument vector
$dist(\mathbf{y}, \hat{\mathbf{y}})$	distance measure between $\mathbf{y}$ and $\hat{\mathbf{y}}$
$e[n]$	linear prediction residual or the excitation signal
$e_p[n]$	the p^{th} subband of $e[n]$
e_{SNR}	error between the original and estimated SNR
λ	sparseness parameter
ϵ	factor determining the noise energy which changes with the desired SNR
ε	convergence criterion for supervised NMF
ϵ_0	small positive value for the weight of the added atom
$f_1[p],\ f_2[p]$	cutoff frequencies for the p^{th} filterbank
f_p	center frequency for the p^{th} band
f_s	sampling frequency
$g[i]$	present GCI candidate
g_{iest}	initial estimated GCIs
g_{cand}	candidate GCIs
$h_B[n]$	impulse response of the bandpass filter
$h_B[f]$	frequency response of the bandpass filter
$h_p[n]$	impulse response of p^{th} filter in the filterbank

Notations

$\mathbf{H}$	Hessian matrix
$KL(\mathbf{y}\|\|\hat{\mathbf{y}})$	Kullback- Leibler divergence between $\mathbf{y}$ and $\hat{\mathbf{y}}$
L_2	Euclidian norm
L_1	Manhattan norm
M_{ns}	total number of noise dictionaries
M_{sp}	total number of speaker dictionaries
N	total number of dictionary atoms
ω_k	resonant frequency for the k^{th} subband signal
$\eta[n]$	randomly chosen +1 or -1 with equal probability
p_B^j	positive peaks between successive zero crossings of s_B^j
p_{B1}^j	subset of p_B^j above T_{P1}^j
p_{B2}^j	subset of p_{B1}^j above T_{P2}^j
$p_r[i]$	reference pitch period
$^\intercal$	transpose
$\phi_k[n]$	phase/frequency modulation component
$\psi()$	Teager energy operator
$r[n]$	room impulse response
$s[n]$	speech utterance
$s_z[n]$	mean removed speech utterance
$s_N[n]$	normalized speech utterance
$s_B[n]$	bandpass filtered signal
s_B^j	bandpass filtered signal between the first and the last zero crossing in the j^{th} frame
$s_r[n]$	reverberant speech signal
SNR^o	original SNR
SNR^e	estimated SNR
SDR^k	SDR with respect to dictionary $\mathbf{D}^k$ for a feature y
SW^k	sum of weights with respect to dictionary $\mathbf{D}^k$ for a feature y
$TSDR^k$	sum of SDR^k over a subset of features
TSW^k	sum of SW^k over a subset of features
T_{P1}^j	first pass positive threshold
T_{P2}^j	second pass positive threshold
T	total number of subbands
T_w	threshold on within-CS
T_b	threshold on between-CS
$v[n]$	vocal tract impulse response
v_{B2}^j	similar to p_{B2}^j obtained from the valleys (negative peaks) between successive zero crossings
$w[n]$	hamming window function
$\mathbf{x}$	weight vector corresponding to $\mathbf{D}$ after source recovery
$\mathbf{x}^k$	weight vector corresponding to $\mathbf{D}^k$ after source recovery
$\mathbf{x}_{sp}^i$	weight vector corresponding to $\mathbf{D}_{sp}^i$ after source recovery
$\mathbf{x}_{ns}^j$	weight vector corresponding to $\mathbf{D}_{ns}^j$ after source recovery
$\mathbf{y}$	original feature vector
$\hat{\mathbf{y}}$	estimated feature vector

$\mathbf{y}_{sp}^{i}$	feature vector corresponding to the i^{th} speech source
$\mathbf{y}_{ns}^{j}$	feature vector corresponding to the j^{th} noise source
$\hat{\mathbf{y}}_{sp}^{i}$	estimated feature vector corresponding to the i^{th} speech source
$\hat{\mathbf{y}}_{ns}^{j}$	estimated feature vector corresponding to the j^{th} noise source
$y[n]$	noisy speech signal
$y_{sp}^{i}[n]$	clean speech component of noisy speech signal for the i^{th} source
$y_{ns}^{j}[n]$	noise component of noisy speech signal for the j^{th} source
$\hat{y}_{sp}^{i}[n]$	estimated speech component of noisy speech signal for the i^{th} source
$\hat{y}_{ns}^{j}[n]$	estimated noise component of noisy speech signal for the j^{th} source

Publications from the Thesis

Conference publications

- V. R. Lakkavalli, K V Vijay Girish, A G Ramakrishnan, "Sub-band Envelope Approach to Obtain Instants of Significant Excitation in Speech," in *Communications (NCC), 2012 National Conference on*, IEEE, 2012
- V. R. Lakkavalli, K. V. Vijay Girish, S. Harshavardhan, A G Ramakrishnan and T V Ananthapadmanabha, "Subband Analysis of Linear Prediction Residual for the Estimation of Glottal Closure Instants," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 945–949, IEEE, 2014
- K V Vijay Girish, A G Ramakrishnan and T V Ananthapadmanabha, "Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation," in *INTERSPEECH*, 2016
- K V Vijay Girish, Veena Vijai and A G Ramakrishnan, "Relationship between spoken Indian languages by clustering of long distance bigram features of speech," in *India Conference (INDICON), 2016 IEEE Annual*, IEEE, 2016
- K V Vijay Girish, T V Ananthapadmanabha and A G Ramakrishnan, "Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources," in *India Conference (INDICON), 2016 IEEE Annual*, IEEE, 2016

Journal submitted

- T V Ananthapadmanabha, K V Vijay Girish and A G Ramakrishnan, "Relative occurrences and difference of extrema for detection of transitions between broad phonetic classes," submitted to *Sadhana, Indian Academy of Sciences*

Technical reports

- T V Ananthapadmanabha, K V Vijay Girish and A G Ramakrishnan, "Detection of transitions between broad phonetic classes in a speech signal," *arXiv preprint arXiv:1411.0370,* 2014

- K V Vijay Girish, A G Ramakrishnan and T V Ananthapadmanabha, "Adaptive dictionary based approach for background noise and speaker classification and subsequent source separation," *arXiv preprint arXiv:1411.0370,* 2016

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An audio signal is a composition of sounds recorded using a microphone. Sounds produced by multiple sources can be active at the same time like many people speaking and music playing in the background. An audio signal conveys the information about the background, the content of the speech sounds if any, the number of audio sources, the relative intensity of each audio source and the variation of the characteristics of sound with time. Speech is any sound produced by human beings to communicate thoughts and feelings, whereas background noise is any sound other than speech like traffic, babble or factory noise. Human beings can easily extract the different information in the audio signals but computer systems are incapable of performing at par with humans, especially in the presence of noise. Audio signals of interest to this thesis are a mixture of foreground speech and background noise, also called as noisy speech. Analysis of the noisy speech is complicated and difficult due to the non-stationary nature of the speech component and change of speaker/background noise over time. Also, the background noise can be stationary or non-stationary and the signal to noise ratio (SNR) may vary with time. It is of importance to improve the clarity and intelligibility of noisy speech (speech enhancement), collect information about the speaker and background environment and understand the content of speech.

1.1 Speech and noise analysis

Speech and noise analysis is the study of noisy speech to know the speaker and the background noise type, the language in which he is speaking, separate the speech and background noise for enhancing the speech and understand the speech content (recognition). Processing of noisy speech signals has been approached by the research community for various independent applications like speaker [1] and noise [2] classification, source separation [3], enhancement [4, 5, 6], speech recognition [7], audio coding [8, 9], prosody modification [10] and speaker normalization [11]. Many of these applications come under the general area of machine listening [12] which extracts useful information from noisy speech signals, and attempts to understand the content akin to human beings. Altering the speech parameters like prosody (duration and pitch period) is useful for voice modification [13] whereas text-

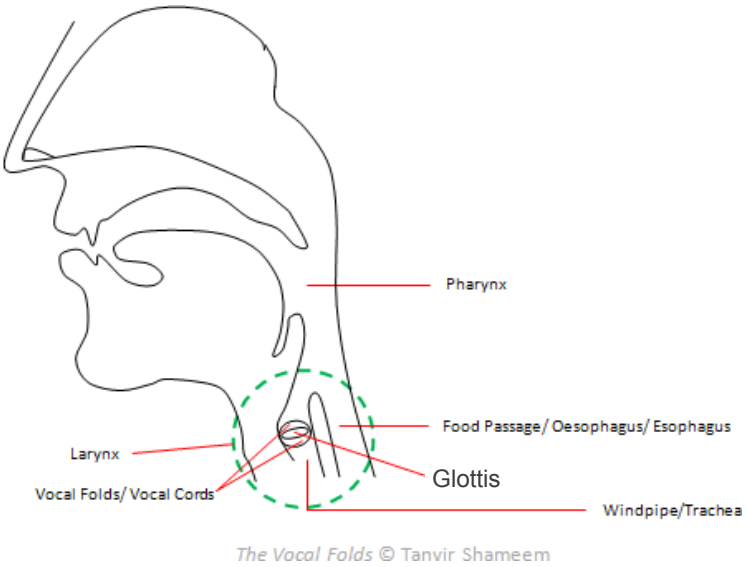
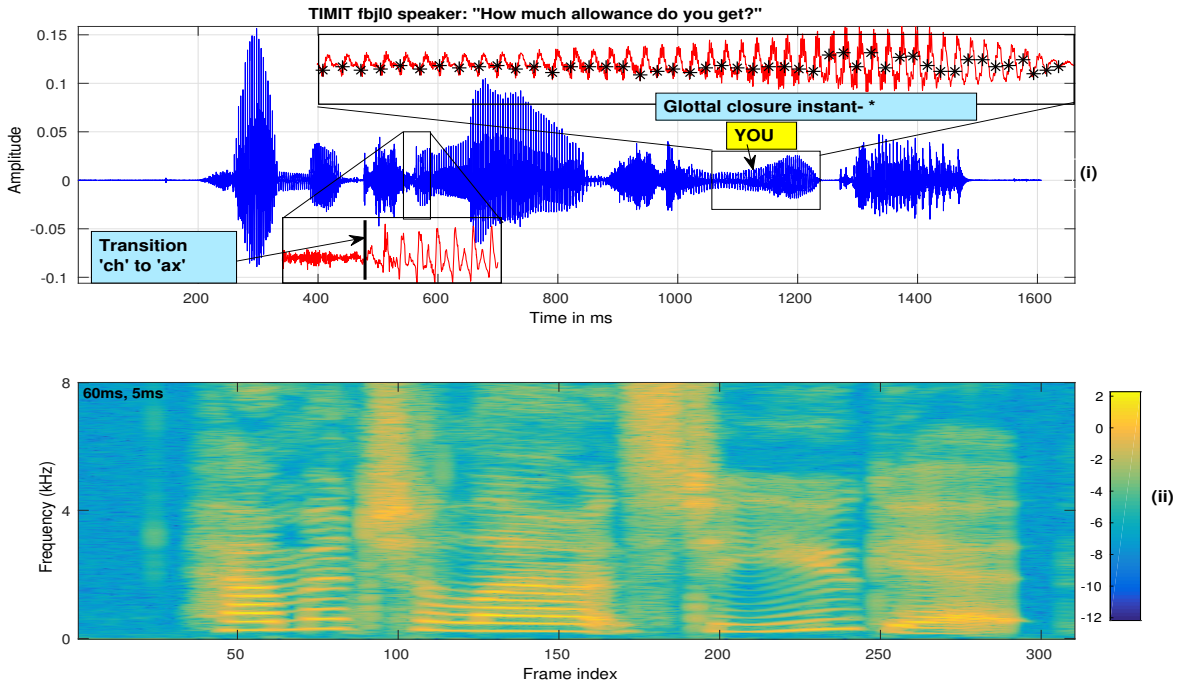


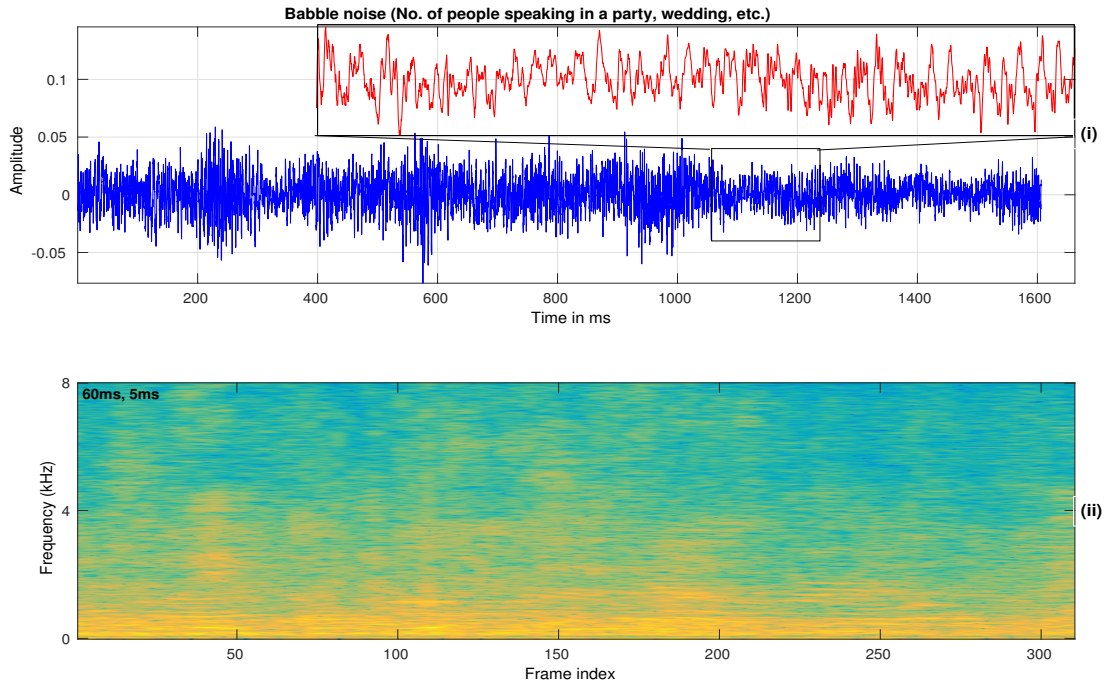
Figure 1.1: Illustration of glottis [15]

to-speech synthesis [14] systems generate speech for automated screen readers. In this thesis, we have addressed various aspects of speech analysis like speaker and noise classification, source separation for speech enhancement, detection of significant transitions and estimation of glottal closure instants (GCIs) useful for speech recognition, segmentation, prosody modification and speech synthesis. GCIs indicate the instants when the glottis (see Fig.1.1) closes abruptly during the vibration of the vocal folds. We also analyze the noise signals so as to classify the same when they occur in noisy speech for better enhancement and separation.

Figure 1.2 shows the time domain signal (i) and the corresponding spectrogram (ii) for (a) a clean speech utterance and (b) a segment of babble noise. The speech utterance is taken from TIMIT [16] female speaker named as *fbjl0* speaking *"How much allowance do you get"*, whose duration is 1.6 seconds. In Fig. 1.2(a)(i), the word *"you"* has been zoomed in to observe the signal (red color) variation in detail. It is seen that the signal is quasi-periodic in nature and we have marked the GCIs detected by our algorithm described in Chapter 4. We can also observe the transition from the unvoiced affricate /ch/ to the vowel /ax/ at 560 ms. The transition segment (red color) has been zoomed in to observe the transition in detail. Phonemes are the units of sound perceptually distinct for the natives of a language. Voiced phonemes [17] are produced due to the vibration of the vocal cord, while unvoiced phonemes do not involve any vocal cord vibration.



(a) A clean speech utterance



(b) A segment of babble noise

Figure 1.2: Speech and babble noise, and the corresponding spectrograms

Figure 1.2 (a) (ii) shows the spectrogram, which is a visual representation of the natural logarithm of the absolute value of the short-time Fourier transforms (STFT) of a sequence of overlapping analysis windows of the speech signal. STFT has been obtained by taking the Fourier transform of the Hanning windowed frames of 60 ms duration, and 5 ms time shift. The colorbar showing the data values corresponding to each color is shown. The harmonics of the pitch can be seen in the voiced frames of *"you"* as horizontal stripes.

Figure 1.2 (b) shows a segment of babble noise [18], which is a mixture of many people speaking at the same time. It is seen from the spectrogram that the energy is concentrated in the low frequency region. This noise signal is added to the speech utterance in (a) at an SNR of 0 dB to get the noisy speech signal shown in Fig. 1.3. It is seen here that it is difficult to accurately detect the transition between /ch/ and /ax/ as compared to the clean speech utterance. Also, the quasi periodicity in the voiced segment *"you"* is corrupted, which makes it difficult to detect the GCIs.

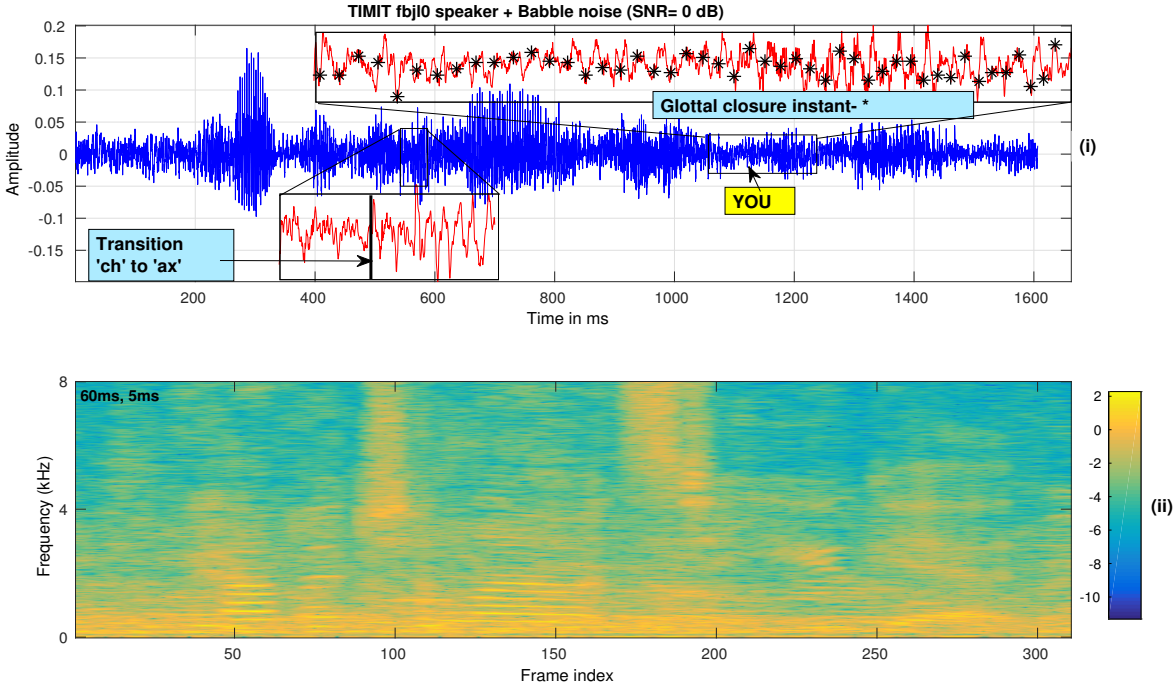


Figure 1.3: Noisy speech signal

In this thesis, we have addressed the problem of speech enhancement and separation using sparse representation based approach, showing how classification of speaker and noise type helps in enhancement. Further, we have used acoustic phonetics knowledge based approach for the detection of transitions and GCIs in clean as well as noisy speech.

1.1.1 Sparse representation

A dictionary is defined as a matrix $\mathbf{D} \in \mathbb{R}^{\mathrm{P}\times\mathrm{N}}$ with $P < N$, containing N column vectors called atoms, denoted as $\mathbf{d}_n, 1 \leq n \leq N$. All the dictionary atoms are normalized to unit L_2 norm. Given a feature vector $\mathbf{y} \in \mathbb{R}^{\mathrm{P}}$ and the dictionary $\mathbf{D}$, the objective is to find the weight vector $\mathbf{x} \in \mathbb{R}^{\mathrm{N}}$, whose elements are the weights corresponding to each dictionary atom, $\mathbf{d}_n$. Then, the equation $\mathbf{y} = \mathbf{D}\mathbf{x}$ either has no solution, if $\mathbf{y}$ is not in the span of the columns of $\mathbf{D}$ or has infinitely many solutions. The problem of having no solution is circumvented by making $\mathbf{D}$ a full rank matrix to span the entire space $\mathbb{R}^{\mathrm{P}}$. It has been shown [19, 20] that a unique solution can be obtained by imposing sparsity on the weight vector $\mathbf{x}$. By constraining the number of non-zero elements in $\mathbf{x}$ to be less than some value dependent on the mutual coherence or spark [19, 20] of the dictionary $\mathbf{D}$, a unique value can be obtained for $\mathbf{x}$. Spark of a dictionary $\mathbf{D}$ is defined as the smallest number of columns from $\mathbf{D}$ that are linearly dependent while the mutual coherence of $\mathbf{D}$, denoted as $\mu(\mathbf{D})$ is defined as the largest of the pairwise absolute inner product between different columns of $\mathbf{D}$:

$$\mu(\mathbf{D}) = \max_{1\leq n,j\leq N, n\neq j} |d_n^T d_j| \tag{1.1}$$

For practical applications, it is desired to get an approximate solution by relaxing the equality as $\mathbf{y} \approx \mathbf{D}\mathbf{x}$. Also, to get a solution which tends to be unique, an upper bound on the sparsity is imposed by posing a constrained optimization problem as:

$$\min_{\mathbf{x}} dist(\mathbf{y}, \mathbf{D}\mathbf{x}) \; subject \; to \; ||\mathbf{x}||_0 \leq m \tag{1.2}$$

where $dist(\mathbf{y}, \mathbf{D}\mathbf{x})$ is the distance measure between $\mathbf{y}$ and $\mathbf{D}\mathbf{x}$ such as L_1, L_2 norm and m is the upper bound on the number of non-zero elements of $\mathbf{x}$.

This method of estimating weights is termed as sparse coding or source recovery. Matching pursuit [21], orthogonal matching pursuit (OMP) [22], basis pursuit [23], focal underdetermined system solver (FOCUSS) [24] and active-set Newton algorithm (ASNA) [25, 26] are some of the source recovery algorithms.

Dictionary learning is a machine learning method to obtain a matrix $\mathbf{D}$, such that the training data is a linear combination of columns (atoms) of the dictionary $\mathbf{D}$, and the corresponding weight vector $\mathbf{x}$ is sparse. Initial work on DL was carried out by Olshausen et al. [27] and Lewicki et al. [28] using probabilistic model of the features. Several methods have been proposed for dictionary learning (DL): random selection of observations [25], K-means clustering [29], vector quantization [30], dictionary update [31], K-SVD [32], simultaneous codeword optimisation (SimCO) [33] and fast dictionary learning [34]. The relation between vector quantization and DL was shown by [30]. DL and source recovery methods have been used for classifying the objects in images by learning class-specific dictionaries [35]. Shafiee et al. [36] have used three different DL methods to classify faces and digits in images.

Sparsity based methods have been used for single channel source separation by Smaragdis et al. [37],

Gao et al. [38] and Badawy et al. [39]. Multi-channel source separation using sparse representation has been addressed by Bofill et al. [40] and Ozerov et al. [41].

1.1.2 Acoustic phonetics

Acoustic phonetics [42] is the study of the acoustic characteristics of speech by analyzing the physical properties of the speech signal and also proposing new representations. Features extracted directly form the time domain speech signal, also known as temporal features, include the short term energy, zero-crossing rate, envelope features and extrema instants. Similar features can also be extracted from the decomposition of the speech signal into various components using methods like subband filtering. The spectral features (frequency based features) obtained by converting the time domain signal into frequency domain are the fundamental frequency, its harmonics, frequency components, spectral centroid, spectral slope, spectral flux etc. Representation of the speech signal using linear prediction (LP) analysis as a convolution of the excitation source and the vocal tract system allows us to model the speech production system. Voiced sounds are produced when the quasi-periodic laryngeal source [42] is filtered by the vocal tract system, while the source for the unvoiced sounds is the turbulent airflow.

Phonemes classified as voiced and unvoiced can further be grouped into different classes based on the manner and place of articulation, which characterize the interactions between the speech organs to produce speech sounds. Table 1.1 shows the various phone groupings based on the TIMIT database [43]. Different phonetic groups have different characteristics; for example, stops are characterized by a sudden explosion of sound or transient preceded by silence (unvoiced stop) or a low energy periodic segment (voiced stop).

1.2 Contributions of the thesis

In this thesis, we first carry out a broad analysis of noisy speech signals using sparse representation for classification and enhancement of speech and noise components. Then, we look deeper into the speech signal to obtain further classification and segmentation into broad phonetic classes by detecting transitions. We also analyze the quasi periodic segments to estimate the GCIs. Thus, various aspects of speech and noise analysis have been addressed in this thesis. Figure 1.3 shows a small segment of noisy speech signal and certain events like transitions and GCIs that we have worked on in this thesis. So, classification of an audio signal is addressed at various levels. At a higher level, noise and speaker classification is addressed and at an acoustic signal level, various classes are segmented, and further segmentation of the voiced segments into quasi-periodic segments is accomplished by estimating GCIs.

Sparse representation based approach has been used to represent the STFT features extracted from the audio signal as a non-negative linear combination of a few dictionary atoms. So, a generative dictionary model is assumed for audio signal representation, using which various measures and algorithms are proposed to classify the noise type and the speaker. It is shown that estimating the noise type and the speaker helps in better speech enhancement and separation using the same dictionary

Table 1.1: Phonetic grouping [43]

Vowels	**Nasals**	**Semivowels/ Glides**	**Fricatives**	**Stops**	**Affricates**	**Silence**
			Voiced	*Voiced*		*Voiced closure*
iy	m	l	z	b	jh	bcl
ih	n	r	zh	d	ch	dcl
eh	ng	w	v	g		gcl
ey	em	y	dh			
ae	en	hh	*Unvoiced*	*Unvoiced*		*Unvoiced closure*
aa	eng	hv	s	p		pcl
aw	nx	el	sh	t		tcl
ay			f	k		kcl
ah			th			*Others*
ao						pau
oy						epi
ow						h#
uh						
uw						
ux						
er						
ax						
ix						
axr						
ax-h						

based representation. Further, the transition between pairs of different noises and speech segments are detected. We have also shown that in the case of an unknown noise/speaker, doing an update of the dictionary using the test signal itself gives a significant improvement in the enhancement.

Knowledge of acoustic phonetics based on observation of the speech signal and its decomposition into subband signals has been used to detect transitions between the phonetic classes for possible speech segmentation. We have obtained transitions with good temporal accuracy for clean speech and acceptable results in the case of noisy speech. LP analysis has been used to estimate the linear prediction residual (LPR), which, being an approximate representation of the source signal, contains good temporal information of the GCI. As the LPR is very noisy, subband decomposition and smoothing using envelope extraction of the subbands is resorted to, in order to narrow down to the approximate region of occurrence of GCIs.

1.3 Organization of the thesis

Every chapter in this thesis begins with an introduction section describing the motivation, literature review and our contributions in that area. It is then followed by the problem statement, proposed method, experimental observations, results and conclusion.

In Chapter 2, we first analyze the dictionaries learnt from noise and speaker sources. Individual dictionaries are learnt for noise and speaker sources to characterize the sources. Then, measures for

frame-wise noise classification are proposed and results are presented. This helps is seeing the confusion arising between different noises. This work on noise classification has been published in [44]. We then address speaker classification in the case of clean speech. Then, we address the problem of noise and speaker classification in a noisy speech signal and observe the performance with varying SNRs, which has been published in [45]. Then, using sparse representation, separation and enhancement of the noisy speech is carried out and the results have been presented in [46]. We have also classified speakers and separated mixed speech in the case of overlapped speech with two speakers talking at the same time.

In Chapter 3, we address the problem of detection of transitions between broad classes in a speech signal. Temporal measures are derived from the speech and its bandpass version to detect transitions and five broad homogeneous classes using a rule and decision tree based classifier. The relation between the defined classes and the phonetic classes is studied and a good temporal accuracy of detection of transitions and distribution of phonetic classes among the five classes is observed. Also, robustness in the presence of Schroeder, white and babble noise is studied [47].

In Chapter 4, we estimate GCIs using subband analysis of the LPR of the speech signal. Bandpass filtered signals capture the harmonics of the fundamental frequency in the LPR and hence the extrema between zero crossings help to narrow down to the GCIs. Smoothing is achieved using piecewise cubic Hermite interpolating polynomial (PCHIP) for envelope extraction. This work on the subband analysis of LPR has been published in [48].

The final chapter concludes the thesis with a summary of the contributions/results and possible directions for further research.

Chapter 2

Sparse representation based classification and separation of noise and speaker sources

A judicious combination of dictionary learning methods, block sparsity and source recovery algorithms are used in a hierarchical manner to classify the noises and the speakers in a noisy speech signal simulated using various combinations of speaker and noise sources, with varied SNR values, down to -10 dB. Fifteen each of randomly chosen male and female speakers from the TIMIT database and all the noise sources from the NOISEX database are used for the simulations. A subset of test features are selected based on maximum correlation measure for noise classification. For speaker classification, a subset of high energy test features are used to recover weights from a concatenated dictionary of subsets of atoms corresponding to high energy features. Speech and noise are separated using dictionaries of the estimated speaker and noise, and an improvement of signal to distortion ratios of around 10 dB is achieved at an SNR of 0 dB. K-medoid and cosine similarity based dictionary learning methods lead to better recognition of the background noise and the speaker. Experiments are also conducted on cases, where either the background noise or the speaker is outside the set of trained dictionaries. In such cases, adaptive dictionary learning leads to performance comparable to the other case of complete dictionaries. In the case of overlapped speech with 2 speakers, we get a speaker classification accuracy of around 84% at a speaker 1 to speaker 2 ratio of 0 dB.

2.1 Introduction

In this chapter, we propose various methods for noise and speaker classification using sparse representation. Then, we devise methods to classify the type of noise and speaker in a noisy speech signal and subsequently separate speech and noise components. Then, we classify speakers from overlapped speech [49] with two speakers mixed at various speaker 1 to speaker 2 ratios in decibel (S1S2R). Dictionaries are learnt from the training data belonging to various speaker/noise sources using a novel

cosine-similarity based dictionary learning method. We develop a system for classification and separation of audio signals. The main objective of this chapter is to show how generative representation of the audio signals as non-negative linear combination of non-negative atoms helps to estimate the speaker/noise sources, which in turn is useful for better separation and speech enhancement. Further, in the case of unknown noise or speaker, the dictionary corresponding to the speaker/noise is updated on the go and the subsequent improvement in separation and classification is observed.

2.1.1 Motivation

Audio signals occurring in nature which are generally of interest to us are mixtures of foreground speech with background noise like factory or babble noise. Analysis of these audio signals is useful for acquiring information about the speaker, background noise environment, location of speech segments and source separation. Identification of background noise can help us to narrow down to possible geographical locations of the speaker while speaker identification helps to reveal his/her identity and selecting an appropriate speaker model for source separation.

The nature of noise in an audio signal varies with the environment such as traffic, restaurant, railway and bus station. Even competing speakers and music may impair intelligibility of speech. Overlapped speech is a case of noisy speech signal where the noise is speech from another speaker. In the case of speech enhancement [5] and audio source separation, especially for hearing impaired [50, 51], the suppression of background audio for improving the intelligibility of speech would be more effective, if the type of background noise can be classified. Other interesting applications of noise identification are forensics [52], machinery noise diagnostics [53], robotic navigation systems [54] and acoustic signature classification of aircrafts or vehicles [55].

Apriori estimates of speaker and the background noise are useful for speech enhancement, separation and speech recognition; which have been of common interest to research community, with many applications in the real world. Frame-wise energy estimation of the separated speech source is useful for identifying speech segments. If this is possible, then the low SNR recordings can be automatically processed to extract only the speech regions. These speech segments can then be processed by human experts, in defense applications, such as analysis of noisy intercepts.

In the case when both speaker and background noise are unknown, the noisy speech can be mapped to the nearest noise and speaker index, and the dictionary for the same can be used for separation or enhancement. The dictionary corresponding to the nearest noise/speaker index can be adapted using the segments of the test signal containing noise/speech only segments. Estimated SNR gives us the temporal information of noise/speech only segments occurring within an audio signal. So, analysis of audio signals can be carried out even if the noise/speaker components belong to an unknown set.

We address the problem of classification and separation of a mixed audio signal containing multiple speakers and noises using the concept of dictionary based representation, block sparsity [56] and sparse non-negative recovery [25]. The advantage of using dictionary based approach for classification is that sparse representation using dictionaries can assume that the signal to be classified may be mixed

with noises whose dictionary is known or can be estimated. We propose a novel, rule-based, sparse representation approach to first identify the type of noises and speakers present in a mixed audio signal and subsequently separate the speech and noise signals.

2.1.2 Literature review

Noise classification can be seen as a first step in machine listening [12], which enables the system to know the background environment. Classification of noise types has been reported in the case of pure noise sources. Kates [57] addressed the problem of noise classification for hearing aid applications based on the variation of signal envelope as feature. Maleh et al. [58] used line spectral frequencies as features for classification of different kinds of noise as well as noise and speech classification. Casey [59] proposed a system to classify twenty different types of sounds using a hidden Markov model classifier and a reduced-dimension log-spectral features. Chu et al. [60] recognized fourteen different environmental sounds using matching pursuit based features combined with mel-frequency cepstral coefficients. Liu et al. [61] devised a TV broadcast video classifier using hidden Markov model (HMM) with audio features. Zhang et al. [62] and Lu et al. [63] segmented and classified audio signals using statistical analysis of simple audio features and a rule-based classifier. Ma et al. [64, 65] and Couvreur et al. [2] devised a HMM based noise classifier for context awareness. Cherla et al. [66] and Ramasubramanian et al. [67] proposed a novel technique for audio analytics and audio indexing using template based modeling of audio classes and HMMs. Ramasubramanian et al. [68] addressed the problem of audio indexing into a target and background class using Gaussian mixture models. Giannoulis et al. [69] conducted a public evaluation challenge on acoustic scene classification (similar to noise classification), where eleven algorithms were evaluated along with a baseline system. The algorithms use time and frequency domain features extracted from the audio signal followed by a statistical model or majority vote based classifier. Cauchi [70] used non-negative matrix factorization for classification of auditory scenes. Techniques for stationary and non-stationary environmental sound recognition was surveyed by Chachada et al. [71]. Malik [72] estimated the amount of reverberation and background noise variance using a statistical technique.

Sparsity based speaker identification using discriminative dictionary learning was done by Tzagkarakis et al. [73] while non-negative matrix factorization for feature extraction was explored by Joder et al. [74]. Representation of audio signals as a sparse, linear combination of non-negative vectors called as dictionary atoms has been used for audio source separation [3, 41, 75], recognition [76, 77], classification [78, 79] and coding [8, 9].

We have used the active-set Newton algorithm (ASNA) [25] and supervised NMF (sup.NMF) [80] algorithm for source recovery in the testing phase. The benefits of these algorithms are that it returns non-negative weights and can handle non-stationary signals like speech. The training phase for the classification problem is dictionary learning (DL) from various speaker/noise sources, where different dictionary atoms encompass the variation in the spectral characteristics.

We have classified the speaker and noise type using sparse representation in the case of a single

speaker and noise [45]. This chapter compares other non-negative dictionary learning methods and separates speech and noise sources. In addition, we deal with unknown speaker and noise sources using adaptive dictionaries.

We deal with the simple problems of noise only and speaker only classification as a prelude to the real problem of noise and speaker classification from a noisy speech signal.

2.1.3 Contributions of this work

The main contributions of this chapter are:

- Dictionary learning by using thresholds on the cosine similarity to ensure distinction between the atoms of the same as well as different source dictionaries
- Proposing two new objective measures, namely, the number of non-zero weights and the sum of weights recovered from ASNA [25] using a concatenation of dictionaries [73], for selecting the most likely audio source from a given set
- Selecting different subsets of segments from the noisy speech signal for noise and speaker classification

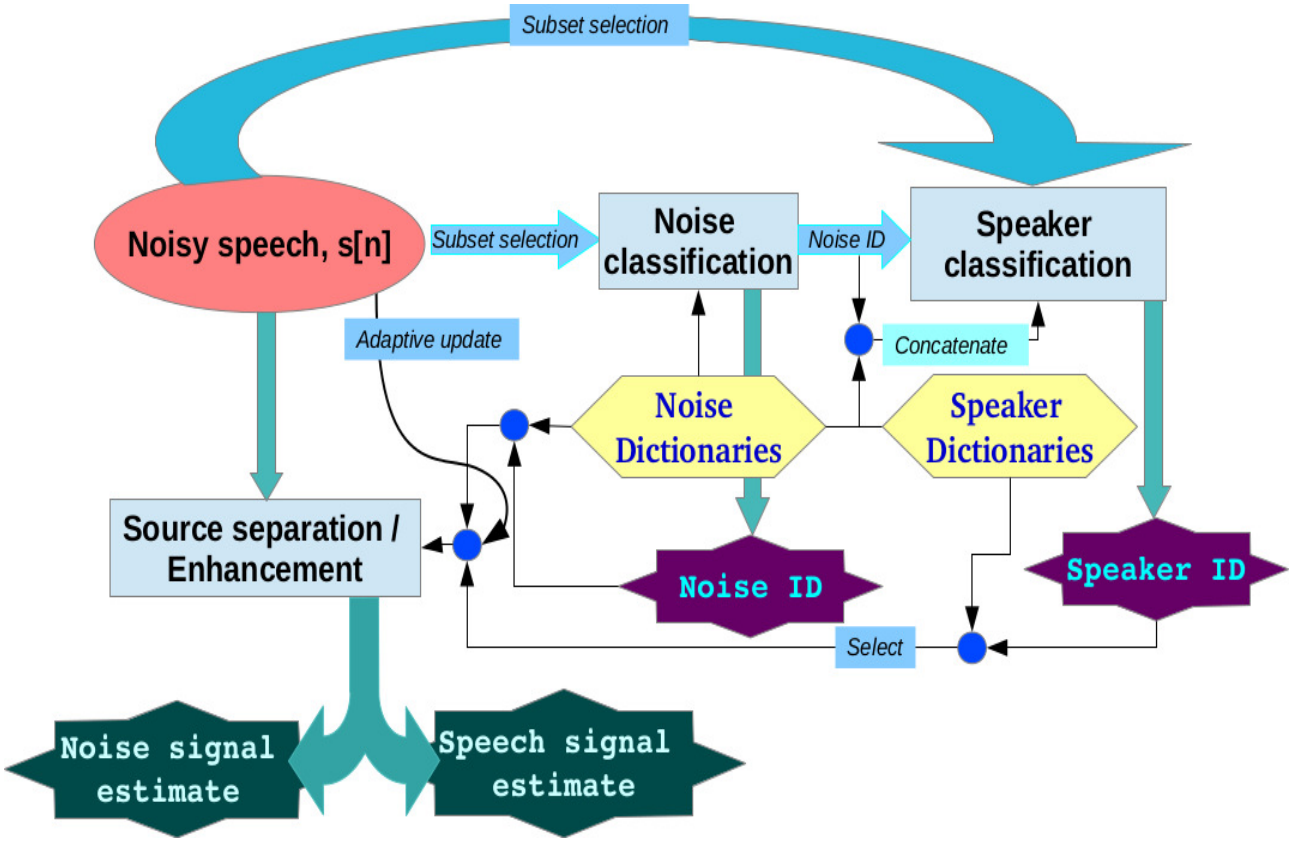


Figure 2.1: Block diagram showing the whole system framework addressed in this chapter

- Block sparsity and concatenated dictionary based classification of multiple speaker and noise sources in a mixed audio signal [81]
- A rule based divide and conquer approach to segment and classify multiple noise segments
- Speaker classification and separation in the case of overlapped speech
- Generalized adaptive update of noise and speaker dictionaries using noise and speech only parts of the noisy speech signal and evaluating the improvement in performance [82]

Figure 2.1 illustrates the framework of the complete system addressed in this chapter. The blue circles in the figure are nodes which receive multiple inputs and feed them to the next block for further processing.

2.2 Dictionary learning

We learn dictionaries for each audio source separately, denoted as $\mathbf{D}^k$ for the k^{th} audio source. All the elements of the dictionary atoms, $\mathbf{d}_j^k$, $1 \leq j \leq N$ are constrained to be non-negative. The constraint is necessary so as to avoid reconstructing mag.STFT features with negative elements and thus to obtain more unique representation of the features. Non-negative matrix factorization (NMF) [80] is a popular technique for learning non-negative dictionaries. Roux et al. [80] reviewed exemplar-based NMF (subset of the training data) and variations of sparse NMF (SNMF) and their application for the speech separation task. In the literature, other non-negative dictionary learning methods have been attempted by Aharon et al. [83] and Bevilacqua et al.[84] as a modification of the K-SVD algorithm [32]. Wang et al. [85] proposed a novel online projected gradient descent to solve the non-negative dictionary learning.

The atoms of the dictionary from the same and different sources are constrained to be as incoherent as possible. Incoherent dictionary learning has been attempted by Ramizez et al. [86] and Barchiesi et al. [87, 88] without any constraint on non-negativity.

Learning discriminative dictionaries is useful for classification by sparse representation approaches. Jiang et al. [89] proposed a label consistent K-SVD algorithm to learn a discriminative dictionary using class labels of training data while Zheng et al. [90] learnt a discriminative dictionary by imposing Fisher discrimination criterion on the sparse coding coefficients.

Our dictionary learning algorithm is non-negative, discriminative and incoherent, the properties necessary for better classification and separation.

2.2.1 Cosine-similarity based dictionary learning

Threshold dependent cosine similarity based dictionary learning (TDCS) is proposed by us in [44]. Feature vectors are L_2 normalized for dictionary learning. Any test feature vector can be represented as an additive, non-negative, linear combination of the dictionary atoms.

In TDCS, each dictionary atom is selected such that it is as uncorrelated as possible to the rest of the atoms belonging to the same as well as other sources. The correlation between a pair of atoms $\mathbf{d}_n, \mathbf{d}_j$ is measured using the cosine similarity as:

$$cs(\mathbf{d}_n, \mathbf{d}_j) = \mathbf{d}_n^T \mathbf{d}_j / (||\mathbf{d}_n||_2 ||\mathbf{d}_j||_2) \tag{2.1}$$

Two types of cosine similarity measures are used: (a) within-class cosine similarity (within-CS) defined as $cs_w(\mathbf{d}_n, \mathbf{d}_j),\ \mathbf{d}_n, \mathbf{d}_j \in \mathbf{D}^k, n \neq j$ where $\mathbf{D}^k$ is the dictionary for a specific source; and (b) between-class cosine similarity (between-CS) defined as $cs_b(\mathbf{d}_n, \mathbf{d}_j),\ \mathbf{d}_n \in \mathbf{D}^k,\ \mathbf{d}_j \in \mathbf{D}^h, k \neq h$.

For each source, the dictionary atoms are learnt such that the cosine similarity between the atoms is below a set threshold, chosen based on the desired performance. Here onwards, the n^{th} dictionary atom corresponding to the k^{th} source is denoted as $\mathbf{d}_n^k$. A randomly selected feature vector, denoted as $\mathbf{y}_l$, is taken as the first atom for the first source, $\mathbf{d}_1^1$. The rest of the atoms are learnt by random selection of the feature vectors (excluding features already selected as atoms): l^{th} feature, $\mathbf{y}_l$, is selected as the n^{th} atom, $\mathbf{d}_n^1$ of dictionary $\mathbf{D}^1$ if the maximum of within-CS, $\max_j cs_w(\mathbf{y}_l, \mathbf{d}_j^1),\ j < n$ (similar to coherence in [20]) is less than a threshold T_w. It is to be noted that maximum of within-CS is same as the mutual coherence in Sec.1.1.1.

The selection of dictionary atoms is stopped once the number of dictionary atoms reaches a pre-decided number N. In case N atoms are not obtained, additional features, which do not satisfy the within-class T_w, are appended in the order of increasing $max\ cs_w$.

To learn dictionaries for the subsequent sources, atoms are learnt using an additional constraint: the feature $\mathbf{y}_t$ from the k^{th} source is selected as the n^{th} atom $\mathbf{d}_n^k$ for the k^{th} dictionary $\mathbf{D}^k$, if maximum of between-CS $max\ cs_b(\mathbf{y}_t, \mathbf{d}_j^h)$, $\mathbf{d}_j^h \in \mathbf{D}^h,\ h < k, 1 \leq j \leq N$ is less than a threshold T_b. We denote the maximum of between-CS as cross coherence.

The threshold T_w ensures that atoms within the same source dictionary are as uncorrelated as possible, while T_b ensures that atoms from different source dictionaries are maximally uncorrelated and discriminable. Lower the values of the thresholds T_w and T_b, greater is the uncorrelatedness between the dictionary atoms.

The TDCS algorithm is summarized in Algorithm 1. For the sake of simplicity, the algorithm does not show the appending of additional dictionary atoms when N atoms could not be obtained.

Algorithm 1 *Threshold dependent cosine similarity*

1: **Initialize:** Dictionary index $k = 1$; $\mathbf{D}^1 = \mathbf{d}_1^1 = \mathbf{y}_l$ (randomly selected feature from the 1^{st} source); Atom index $n = 2$; set T_w and T_b as the thresholds.
2: **repeat**
3: Extract L number of mag.STFT features from the k^{th} source denoted as $\mathbf{y}_l, 1 \leq l \leq L$.
4: **repeat**
5: If $n > 1$, find the maximum of within-CS, m_w as:

$\max(cs_w(\mathbf{y}_l, \mathbf{d}_j^k)\ \forall\, j = 1...n-1)$

If $k > 1$, find the maximum of between-CS, m_b as:

$\max(cs_b(\mathbf{y}_l, \mathbf{d}_j^h)\ \forall\, j = 1..N, h < k)$

if $m_w \leq T_w$ and $m_b \leq T_b$ (for $k > 1$) **then**

Assign randomly selected $\mathbf{y}_l$ as the n^{th} atom: $\mathbf{d}_n^k = \mathbf{y}_l$ and append to the dictionary: $\mathbf{D}^k = [\mathbf{D}^k\ \mathbf{d}_n^k]$

$n = n + 1$

end if

until $n > N$

$k = k + 1$; $n = 1$

until All source dictionaries are learnt

end

2.3 Algorithms for source recovery

We have used the ASNA [25, 26] and sup.NMF [80] algorithms for source recovery to efficiently obtain non-negative sparse representations of the test audio signals. Both the algorithms are based on minimizing the generalized Kullback-Leibler (KL) divergence between an observed magnitude spectrum and a non-negative linear combination of atoms.

The minimization problem for supervised NMF [80] can be posed as:

$$\underset{\mathbf{X}}{\text{minimize}}\ KL(\mathbf{Y}||\mathbf{DX}) + \mu||\mathbf{X}||_1\ s.t.\ \mathbf{X} \geq 0 \tag{2.2}$$

$\mathbf{Y}$ is the matrix whose columns are the test features $\mathbf{y}$, $\mathbf{D}$ is the dictionary, $\mathbf{X}$ is the weight matrix and $KL(\mathbf{Y}||\mathbf{DX})$ is the KL divergence between $\mathbf{Y}$ and $\mathbf{DX}$,

$\mathbf{X}$ is estimated by multiplicative updates [80] using $\mu = 5$ as

$$\mathbf{X} = \mathbf{X} \otimes \frac{\mathbf{D}^\intercal(\frac{\mathbf{Y}}{\mathbf{DX}})}{\mathbf{D}^\intercal + \mu} \tag{2.3}$$

where $\otimes$ denotes the element-wise multiplication and the quotient line is element-wise division. The stopping criterion for the above multiplicative update is when the relative error in the KL divergence falls below $\varepsilon = 0.001$ or the number of iterations exceeds 100.

The minimization problem for ASNA [25] can be posed as:

$$\underset{\mathbf{x}}{\text{minimize}}\ KL(\mathbf{y}||\hat{\mathbf{y}}),\ \hat{\mathbf{y}} = \mathbf{Dx}\ s.t.\ \mathbf{x} \geq 0 \tag{2.4}$$

where $KL(\mathbf{y}||\hat{\mathbf{y}})$ is the KL divergence between $\mathbf{y}$ and $\hat{\mathbf{y}}$, $\mathbf{D}$ is the dictionary and $\mathbf{x}$ is the weight vector.

An extension of the ASNA algorithm is proposed in [26] by adding a L_1 regularization term to

Equation 2.4 (denoted as ASNA-L1) as,

$$\underset{\mathbf{x}}{\text{minimize}}\ KL(\mathbf{y}||\mathbf{Dx}) + \lambda||\mathbf{x}||_1,\ s.t.\ \mathbf{x} \geq 0 \tag{2.5}$$

$||\mathbf{x}||_1$ is the L_1 norm of the weight vector and λ is the sparseness parameter. Here, sparsity constraints are explicitly introduced by adding the L_1 regularization term.

The steps in ASNA-L1 algorithm with the regularization term are listed in Algorithm 2. Given the known dictionary $\mathbf{D} = [\mathbf{d}_1..\mathbf{d}_i..\mathbf{d}_N]$ and the observation vector $\mathbf{y}$, Algorithm 2 estimates $\mathbf{x}$ as a solution to Equation 2.5. As seen in the algorithm, the first active atom is added to the active set $\mathbf{A}$. We then iteratively find the most promising atom not in the active set every J^{th} iteration, and add it to the active set. The weight of the added atom is initialized to a small positive value ϵ_0. Weights for the active atoms are iteratively updated using Newton's method, ensuring that all weights remain non-negative. Atoms whose weights go to zero are removed from the active set. The procedure is iterated until a stopping criterion is achieved [25, 26]. In our work, we fix the total number of iterations as the stopping criterion decided by varying the number of iterations and choosing the best value as shown in Sec.2.5.2.

The convergence rate of ASNA algorithm is expected to be linear whose detailed analysis can be found in [25]. The convergence of sup.NMF depends on the initialization of $\mathbf{X}$ in Equation 2.2 [91].

Algorithm 2 *Active-set Newton algorithm with L_1 regularization [26]*

Choose a value for J ($J = 2$)
Initialize the active set as $\mathbf{A} = []$ ($\mathbf{A}$ contains the set of indices), $\#iteration = 0$
Normalize each dictionary atom $\mathbf{d}_i$ to unity norm
Calculate all the weights using $\mathbf{x}(i) = \dfrac{\mathbf{1}^T\mathbf{y}}{\mathbf{1}^T\mathbf{d}_i + \lambda}$
Select the first active atom using $k = \underset{i}{arg\,min}\ KL(\mathbf{y}, \mathbf{x}(i)\mathbf{d}_i) + \lambda\mathbf{x}(i)$ and add to the active set as $\mathbf{A} = [\mathbf{A}\ k]$
repeat
    Update $\hat{\mathbf{y}} = \sum \mathbf{x}(i)\mathbf{d}_i,\ i \in \mathbf{A}$
    Update Active set
    if $\#iteration \mod J = 0$ **then**
        Find the derivative (gradient) of the KL divergence with respect to weights not in the active set as $\mathbf{grad}_j = \dfrac{d}{d\mathbf{x}(j)}KL(\mathbf{y}, \hat{\mathbf{y}}) = \mathbf{d}_j^T\left(1 - \dfrac{\mathbf{y}}{\hat{\mathbf{y}}}\right) + \lambda, \forall j \notin A$, where $\dfrac{\mathbf{y}}{\hat{\mathbf{y}}}$ is elementwise division
        If the smallest derivative in the above step is negative, add the corresponding atom to the active set as $\mathbf{A} = [\mathbf{A}\ k]$ $s.t.\ k = \arg\min_j \mathbf{grad}_j$ and set its weight $\mathbf{x}(k) = \epsilon_0$
    end if
    Update weights

14: Find the gradient of the KL-divergence with respect to the weights of the active set as $\mathbf{grad_A} = \mathbf{D_A^T}\left(1 - \frac{\mathbf{y}}{\hat{\mathbf{y}}}\right) + \lambda$

15: Find the Hessian matrix as $\mathbf{H_A} = \mathbf{D_A^T} diag\left(\frac{\mathbf{y}}{\hat{\mathbf{y}}^2}\right)\mathbf{D_A}$

16: Update the weights as $\mathbf{x_A} = \mathbf{x_A} - \alpha\mathbf{p}$ where $\mathbf{p}$ is the search direction, $p = (\mathbf{H_A})^{-1}\mathbf{grad_A}$. α is chosen such that weights, $\mathbf{x_A}$ are non-negative

17: Set $\#iteration = \#iteration + 1$

18: **until** Stopping criterion

end

2.4 Analysis of noise/speaker dictionaries

We learn dictionaries for all the fifteen noises from the NOISEX database and thirty speakers from TIMIT database. In this section, we analyze the dictionaries learnt from the noise and speaker sources. Table 2.1 shows the list of 15 noise sources taken from the NOISEX database [18], mainly consisting of military noises and other commonly occurring noises. Each noise is of a duration of 3-4 minutes and 60% is used for training, 20% for validation and the rest 20% for testing. The database for speech sources is taken from randomly selected 15 male and 15 female speakers from dialect 5 of the training set of the TIMIT database [43] as shown in Table 2.2. For each speaker, 8 utterances are used for training and the rest 2, for validation and testing respectively. The duration of each utterance is 2-4 seconds. The validation sets of speaker and noise sources have been used for analysis, tuning the parameters and noise/speaker classification in the case of clean speech/noise signals. Test set has been used for simulating overlapped and noisy speech signals, and getting results on source separation.

Table 2.1: List of the fifteen noise sources from NOISEX [18] database

Sl. no	Noise type	Description
1	'babble'	Speech babble
2	'buccaneer1'	Jet cockpit noise 1
3	'buccaneer2'	Jet cockpit noise 2
4	'destroyerengine'	Destroyer engine room noise
5	'destroyerops'	Destroyer operations room noise
6	'f16'	F-16 cockpit noise
7	'factory1'	Factory floor noise 1
8	'factory2'	Factory floor noise 2
9	'hfchannel'	HF radio channel noise
10	'leopard'	Military vehicle noise
11	'm109'	Tank noise
12	'machinegun'	Machine gun noise
13	'pink'	Pink noise
14	'volvo'	Car interior noise
15	'white'	White noise

Table 2.2: List of the thirty speaker sources used from the TIMIT database [43]

Sl.no	Female	Male
1	fsjg0	mwac0
2	ftlg0	mges0
3	fsms1	mwem0
4	fsmm0	mtdp0
5	fear0	mclm0
6	fpmy0	mjpg0
7	fskp0	mrav0
8	fkkh0	msem1
9	fdmy0	mdwh0
10	fjxm0	mrld0
11	ftbw0	mhmg0
12	fgmb0	mjfh0
13	fgdp0	msdh0
14	flmk0	mdhl0
15	fbmh0	mram0

The following five dictionary learning methods have been used :

1. *Random selection of features*: Features randomly picked up from the training set using a uniform distribution are assigned as the dictionary atoms.

2. *K-medoid clustering* : This is performed using the algorithm proposed by Park et al. [92] and the K medoids are used as the dictionary atoms.

3. *TDCS-0.9*: TDCS algorithm, proposed in Sec. 2.2.1 with the within and between-class thresholds as $T_w = 0.9$, $T_b = 0.9$. Two variants are considered: (a) TDCS-0.9, same as Algorithm 1 and (b) Variable TDCS-0.9, the number of atoms vary, since the threshold is not relaxed.

4. *TDCS-0.8*: TDCS algorithm with the within and between-class thresholds as $T_w = 0.8$, $T_b = 0.8$. Again, two variants are considered: (a) TDCS-0.8, same as Algorithm 1 and (b) Variable TDCS-0.8, by not relaxing the threshold.

5. *SNMF*: Sparse NMF dictionary learnt using SNMF algorithm in [80].

We have chosen to use K-medoids clustering [92] instead of the well-known K-means clustering for two reasons:- (1) the medoids themselves are representative features from the training and avoids averaging of the features and (2) K-medoids are less sensitive to outliers. Feature extraction for dictionary learning is described in the next section.

2.4.1 Extraction of STFT features

Frames of 60 ms duration are extracted with a shift of 15 ms from the training sets of speaker and noise sources. We experimented with different choices of frame size/shift and arrived at these values

as the best. Higher frame size of 60 ms is found to give better classification accuracy and separation performance due to the higher frequency resolution. Features are extracted as the magnitude of the short-time Fourier transform (mag.STFT) of these frames using a Hanning window. The dimension of the feature vector is 481 for a 16 kHz sampling frequency of the speech/noise signal. For dictionary learning, features having very low energy (0.001 times) relative to the average energy of the features are removed. It is to be noted that all the features and dictionary atoms are non-negative as we require a non-negative representation during classification and separation stage. For classification of sources, the features having very low relative energy have been ignored throughout this chapter. Dictionaries for all the sources are learnt separately by the five dictionary learning methods listed above using $N(= 500)$ number of atoms. Features for each speaker and noise source are extracted separately and the corresponding dictionaries are built. All the atoms of the dictionaries are normalized to unit L_2 norm. The dictionaries for M_{sp} speaker and M_{ns} noise sources are denoted as $\mathbf{D}_{sp}^{i}$, $1 \leq i \leq M_{sp}$ and $\mathbf{D}_{ns}^{j}$, $1 \leq j \leq M_{ns}$, respectively

2.4.2 Analysis of the coherence of the dictionaries

Dictionaries are learnt using mag.STFT features and neglecting very low energy features as mentioned in the previous section. In the case of Variable TDCS-0.9 and Variable TDCS-0.8, we see in Fig.2.2 the effect of terminating the algorithm based on the threshold T_w, T_b even though 500 atoms are not obtained. It is observed that four of the noise dictionaries learnt have less than 500 atoms when we do not relax the threshold of 0.9. For machinegun and volvo noise, the number of atoms selected is very low, namely 67 and 54, respectively. For $T_w = T_b = 0.8$, all the noise sources learnt have less than 500 atoms and 9 of the 15 noise source dictionaries have less than 100 atoms. Babble, buccaneer1 and white noise dictionaries have more than 300 atoms, showing the variability in the spectral characteristics of these noise sources. The observations are in line with the inherent structure of the noise sources as babble being a mixture of speech from different people has a very high spectral variability and hence the number of atoms selected is the highest. Machinegun noise is silence followed by impulsive firing of bullets and hence has the lowest spectral variation, neglecting the silence regions. By not relaxing the threshold T_w, T_b, we ensure that the mutual coherence is below 0.9 and 0.8 for Variable TDCS-0.9 and Variable TDCS-0.8 as seen in Fig. 2.3. In case we learn fixed number of dictionary atoms by relaxing the constraints on T_w, T_b as in Algorithm 1, we have to compromise on the mutual coherence as shown in Fig. 2.3 as TDCS-0.9 and TDCS-0.8. So, there is a tradeoff between the number of dictionary atoms and the mutual coherence. It is seen that dictionaries learnt using random selection of features have high mutual coherence as compared to TDCS.

Figure 2.4 shows the number of atoms selected based on the Variable TDCS-0.9 and Variable TDCS-0.8 methods for thirty speakers while Fig. 2.5 shows the mutual coherence of dictionaries learnt using different methods. We observe that the No. of atoms learnt using Variable TDCS-0.8 is less than 200 for 27 of the 30 speaker sources. It is seen that mutual coherence using random selection of features is close to 1.

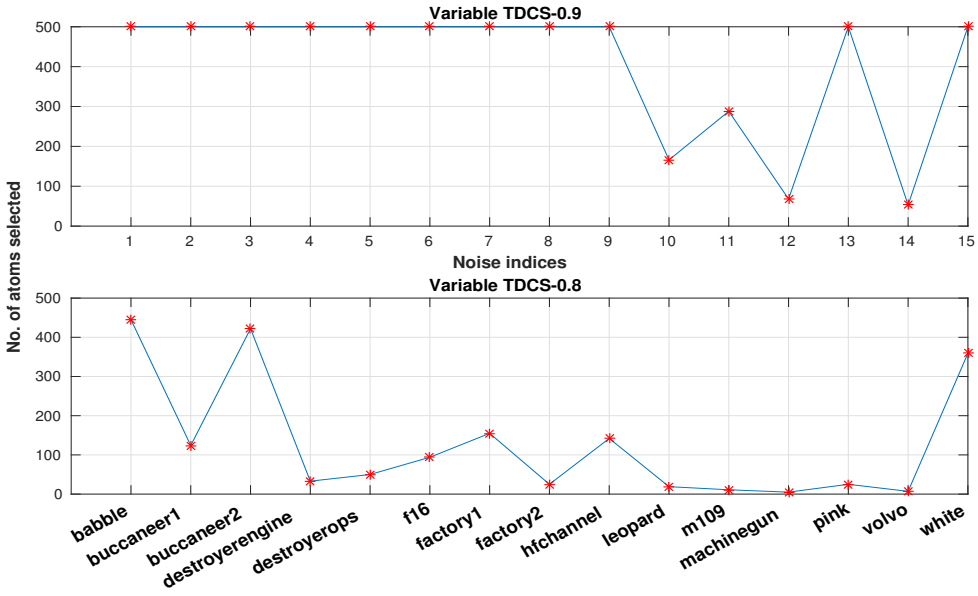


Figure 2.2: Number of atoms selected for different noise dictionaries by not relaxing the threshold

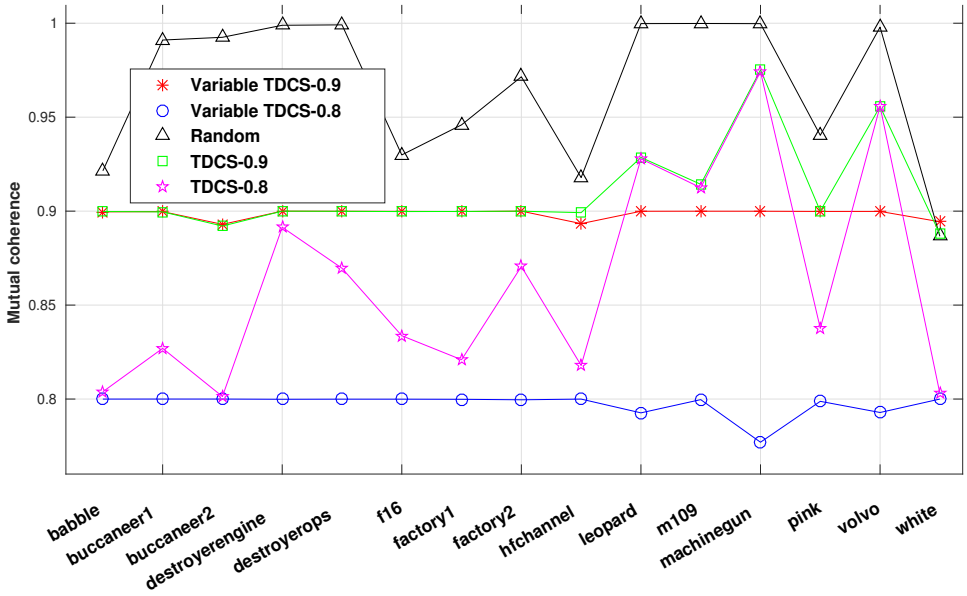


Figure 2.3: Mutual coherence for each individual noise dictionary

Figure 2.6 shows the percentage distribution of number of dictionary atom combinations as a function of within-class cosine similarity and between-class cosine similarity. It is seen that SNMF is skewed towards very high values of within/between class cosine similarity.

2.5 Signal and noise source separation

Dictionary learning and sparse representation have been used for source separation [3, 41, 75], which give the estimates of the separated speech and noise signals from a noisy speech signal. Given the

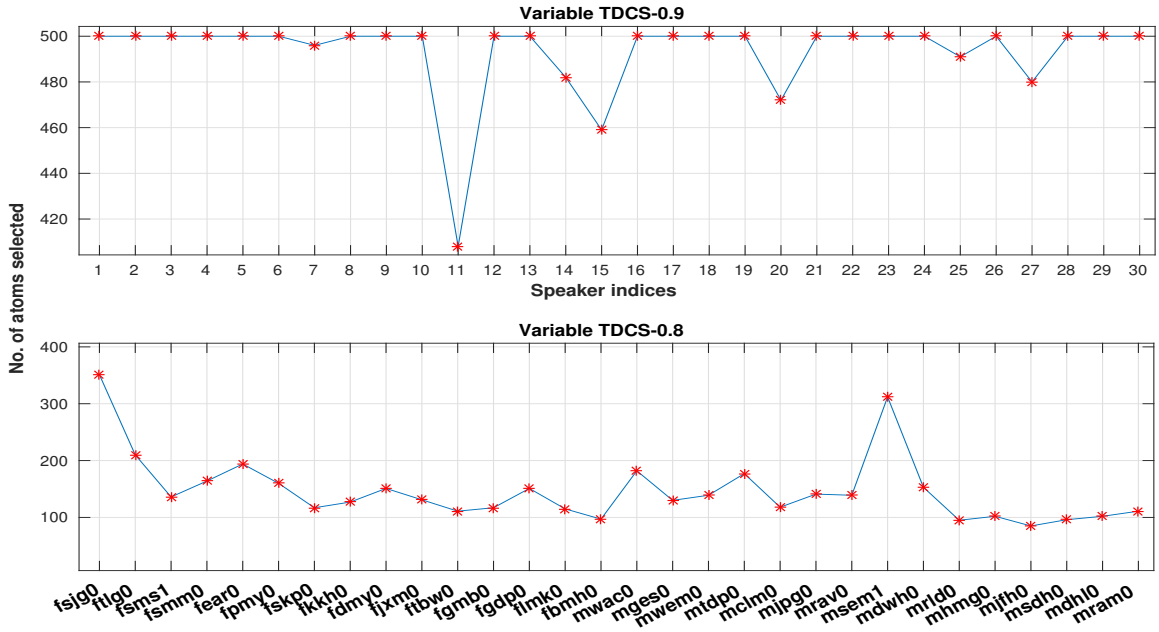


Figure 2.4: No. of atoms selected for speaker dictionaries

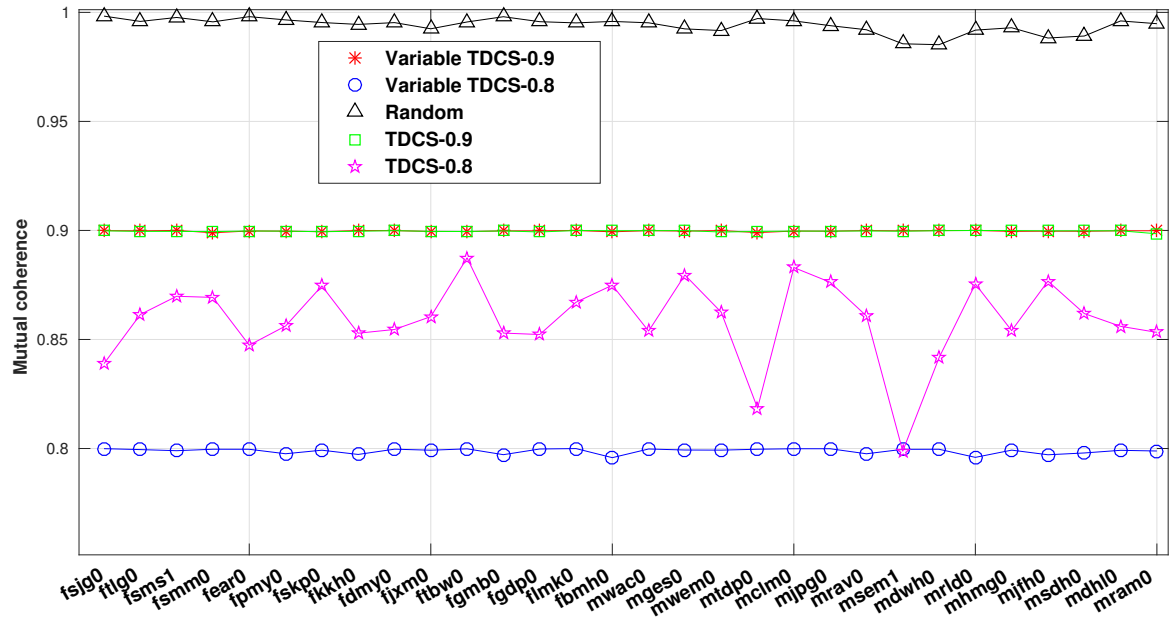
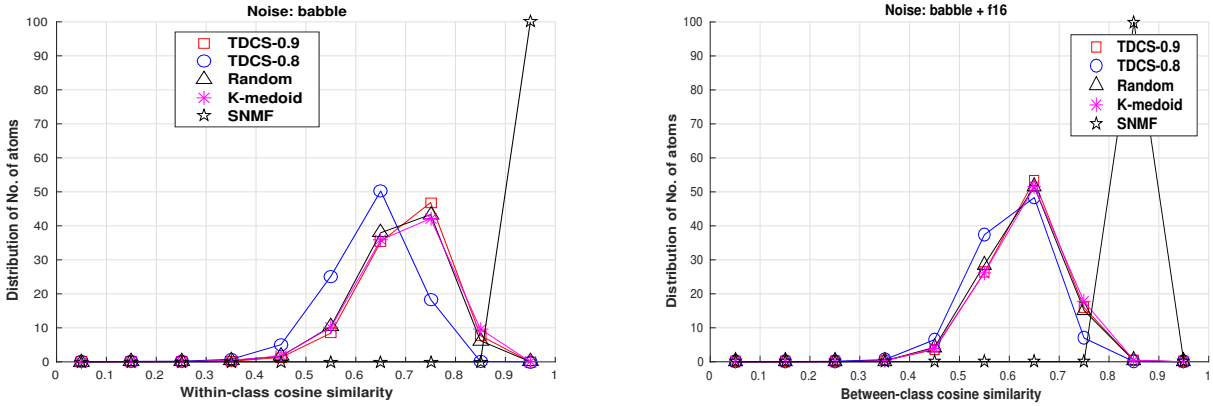
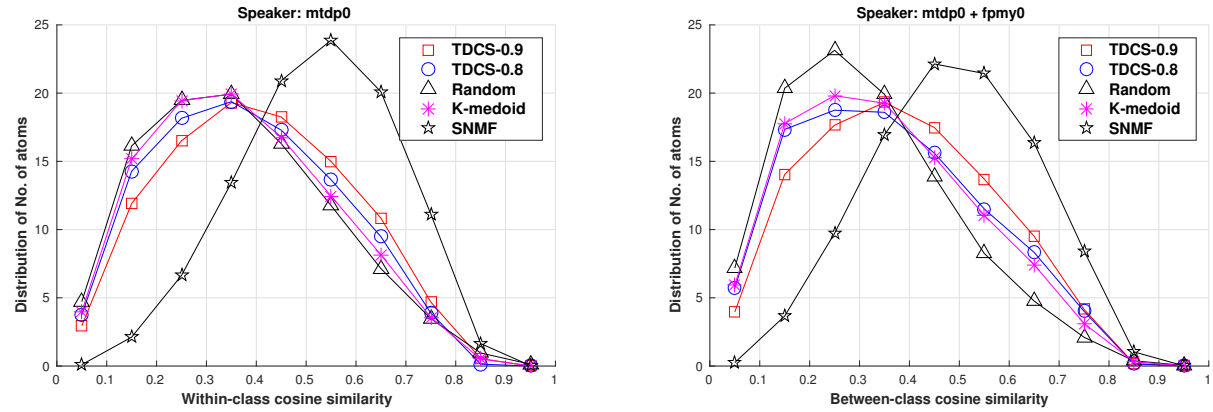


Figure 2.5: Mutual coherence for each individual speaker dictionary

noisy speech feature, $\mathbf{y}$ (mag.STFT feature extracted from the noisy signal $y[n] = y^i_{sp}[n] + y^j_{ns}[n]$ as explained in Sec.2.4.1, $y^i_{sp}[n], y^j_{ns}[n]$ are the original speech and noise signals) which can be seen as a linear combination of the i^{th} speech source, $\mathbf{y}^i_{sp}$ and the j^{th} noise source, $\mathbf{y}^j_{ns}$ as $\mathbf{y} = \mathbf{y}^i_{sp} + \mathbf{y}^j_{ns}$



(a) Noises



(b) Speakers

Figure 2.6: Percentage distribution of dictionary atom pairs as a function of within-class and between-class cosine similarity for (a) two noises; (b) two speakers

(neglecting the phase component of the STFT feature), we need to estimate the speech and noise components as $\hat{\mathbf{y}}^i_{sp}$ and $\hat{\mathbf{y}}^j_{ns}$. The separation problem can be seen as:

$$\mathbf{y} \approx \hat{\mathbf{y}}^i_{sp} + \hat{\mathbf{y}}^j_{ns} = \mathbf{D}^i_{sp}\mathbf{x}^i_{sp} + \mathbf{D}^j_{ns}\mathbf{x}^j_{ns} = \mathbf{D}\mathbf{x} \tag{2.6}$$

where $\hat{\mathbf{y}}^i_{sp} = \mathbf{D}^i_{sp}\mathbf{x}^i_{sp}$, $\hat{\mathbf{y}}^j_{ns} = \mathbf{D}^j_{ns}\mathbf{x}^j_{ns}$, $1 \leq i \leq M_{sp}$, $1 \leq j \leq M_{ns}$, $\mathbf{D}^i_{sp}$ and $\mathbf{D}^j_{ns}$ are the known speaker and noise dictionaries for M_{sp} speaker and M_{ns} noise sources, $\mathbf{D}$ is the concatenated dictionary, $\mathbf{D} = [\mathbf{D}^i_{sp}\ \mathbf{D}^j_{ns}]$, and $\mathbf{x}^i_{sp}$ and $\mathbf{x}^j_{ns}$ are the corresponding weights to be estimated. The above equation can be solved by minimizing the KL-divergence between $\mathbf{y}$ and $\mathbf{D}\mathbf{x}$ similar to Equations 2.4,2.5,2.2 which estimate $\mathbf{x} = [\mathbf{x}^{i\intercal}_{sp}\ \mathbf{x}^{j\intercal}_{ns}]^\intercal$ using the recovery algorithm ASNA, ASNA-L1 and sup.NMF, respectively. The speech and noise features $\hat{\mathbf{y}}_{sp}$ and $\hat{\mathbf{y}}_{ns}$ are estimated as

$$\hat{\mathbf{y}} = [\mathbf{D}^i_{sp}\ \mathbf{D}^j_{ns}][\mathbf{x}^{i\intercal}_{sp}\ \mathbf{x}^{j\intercal}_{ns}]^\intercal \tag{2.7}$$

$$\hat{\mathbf{y}}^i_{sp} = \mathbf{D}^i_{sp}\mathbf{x}^i_{sp},\ \hat{\mathbf{y}}^j_{ns} = \mathbf{D}^j_{ns}\mathbf{x}^j_{ns} \tag{2.8}$$

The feature matrix corresponding to the speech and noise is reconstructed from the estimated speech and noise features by normalizing to ensure that the source estimates sum to the mixed features as:

$$\hat{\mathbf{Y}}^i_{sprec} = \mathbf{Y} \otimes \frac{\hat{\mathbf{Y}}^i_{sp}}{\hat{\mathbf{Y}}^i_{sp} + \hat{\mathbf{Y}}^j_{ns}} \tag{2.9}$$

where $\otimes$ denotes the element-wise multiplication and the quotient line is element-wise division. $\mathbf{Y}$ is the matrix whose columns are the noise speech features $\mathbf{y}$, $\hat{\mathbf{Y}}^i_{sp}$ and $\hat{\mathbf{Y}}^j_{ns}$ are the matrices of the estimated speech and noise features $\hat{\mathbf{y}}^i_{sp}$, $\hat{\mathbf{y}}^j_{ns}$. The speech component in the time domain $y^i_{sp}[n]$ is recovered using the reconstructed speech feature matrix, $\hat{\mathbf{Y}}^i_{sprec}$ and phase of the noisy speech signal using overlap and add method as $\hat{y}^i_{sp}[n]$. The corresponding noise signal is estimated as $\hat{y}^j_{ns}[n] = y[n] - \hat{y}^i_{sp}[n]$.

2.5.1 Measures for quantifying separation performance

The following measures are used to evaluate the performance of speech and noise separation:

- *Speech to distortion ratio* (SDR) [93] between the original and the estimated speech signal for the i^{th} source is defined as:

$$SDR^i = 20\log_{10}\frac{||y^i_{sp}[n]||_2}{||y^i_{sp}[n] - \hat{y}^i_{sp}[n]||_2} \tag{2.10}$$

  which is the ratio of the L_2 norms of the original speech signal to the distortion between the original and the estimated speech signal. SDR quantifies the deviation of the estimated speech signal from the original speech signal; higher the SDR, better is the separation performance.

- *Noise to distortion ratio* (NDR) between the original and the estimated noise signal for the j^{th} source is defined as:

$$NDR^j = 20\log_{10}\frac{||y^j_{ns}[n]||_2}{||y^j_{ns}[n] - \hat{y}^j_{ns}[n]||_2} \tag{2.11}$$

- *Error in SNR*: Estimated SNR is defined as the ratio of the L_2 norm of the estimated speech to noise signal in decibel as:

$$SNR^e = 20\log_{10}\frac{||\hat{y}^i_{sp}[n]||_2}{||\hat{y}^j_{ns}[n]||_2} \tag{2.12}$$

while the original SNR uses the ground truth speech and noise signal:

$$SNR^o = 20\log_{10}\frac{||y^i_{sp}[n]||_2}{||y^j_{ns}[n]||_2} \tag{2.13}$$

The error in the estimate of SNR is given by $e_{SNR} = SNR^o - SNR^e$. Mean of the absolute value of the e_{SNR}, denoted as MAE-SNR and the standard deviation of the e_{SNR}, denoted as STD-SNR are used as the measures to quantify the performance of the algorithm.

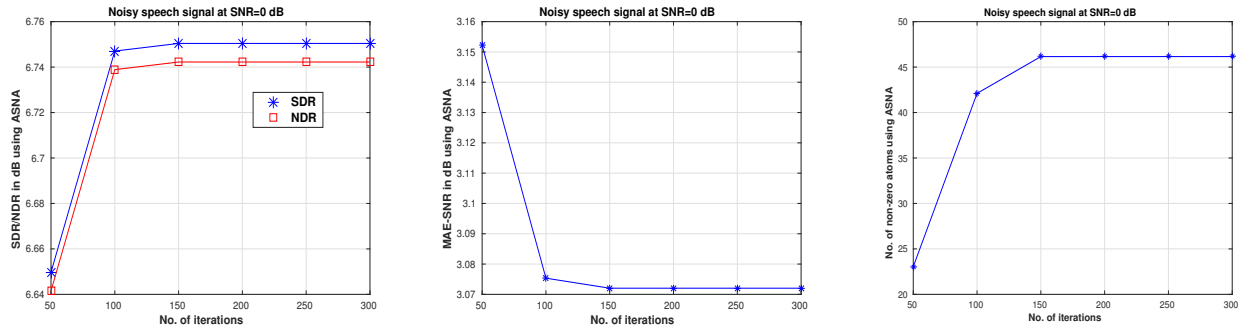
It is to be noted that the SDR and the error in estimated SNR are computed only for the regions/frames, where speech is present.

2.5.2 Variation of separation performance

To find the optimal number of iterations for ASNA and ASNA-L1 algorithms, we use a small development set to evaluate the source separation performance on noisy speech signal. The validation utterances of 'fsjg0' and 'mram0' speakers are mixed with equal duration validation segments of white and babble noise at an SNR of 0 dB for evaluation. The K-medoid dictionaries learnt for the noise and speaker sources are used. We compare the performance of ASNA and ASNA-L1 source recovery methods.

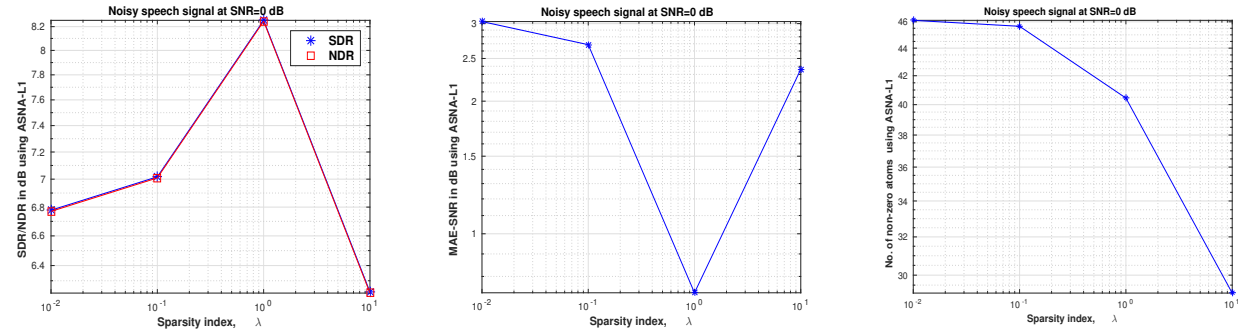
We evaluate the source separation performance using the measures SDR, NDR and estimated SNR for the above four combinations of speaker and noise sources. Figure 2.7 shows the variation of SDR, NDR, estimated SNR and the number of atoms having non-zero weights averaged over all the combinations of noisy speech signals. Number of non-zero weights show how sparse the weight vectors are, averaged over all the features. We arrive at the optimum value of λ for ASNA-L1 by observing the variation of source separation performance on the validation set and choosing the λ which gives the best performance as done in [26]. We vary the value of λ for ASNA-L1 and plotted the same performance measures as shown in Figure 2.8. It is seen that we get maximum SDR and NDR at $\lambda = 1$. In Fig. 2.9, we fix $\lambda = 1$ and see the variation of the performance measures and No. of non-zero weights as a function of the number of iterations. It is observed that the performance saturates at 150 iterations for ASNA and ASNA-L1 at $\lambda = 1$. Also, it is seen that the average number of non-zero weights is around 46 using ASNA while it is around 40 using ASNA-L1 at $\lambda = 1$ which shows that L_1 regularization encourages sparsity. Also, ASNA-L1 gives better SDR of 8.25 dB at $\lambda = 1$ as compared to 6.75 dB using ASNA algorithm. Based on the above analysis, we have used ASNA-L1 with $\lambda = 1$ and the No. of iterations as 150 for source separation.

We experimented using SNMF dictionary to evaluate the source separation performance for the same development set. The weights are estimated using ASNA, ASNA-L1 and sup.NMF (the weights estimated using the update equation for the activation matrix in [80]). It is observed that SDR saturates at 4.46 dB and 7.30 dB using SNMF dictionary with ASNA and ASNA-L1 with $\lambda = 10$. It is to be noted that $\lambda = 1$ gives the best results using K-medoid while $\lambda = 10$ gives the best results using SNMF dictionaries. Table 2.3 compares the source separation performance with respect to SDR,



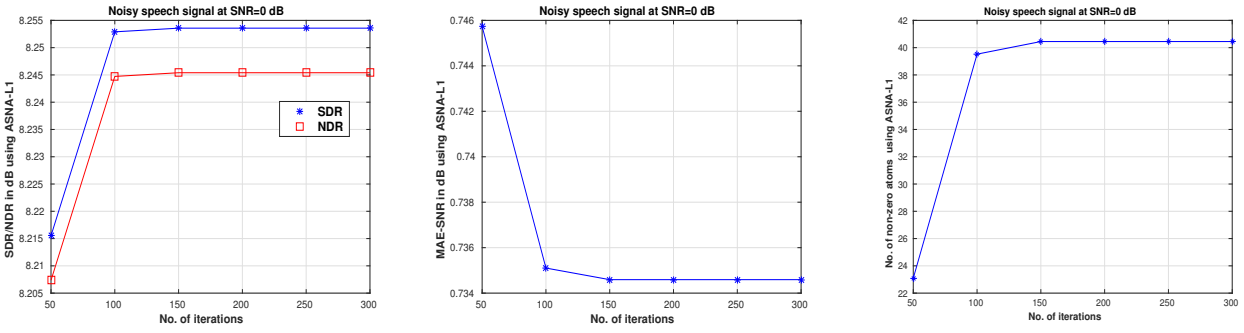
(a) (b) (c)

Figure 2.7: Variation of SDR, NDR, MAE-SNR and No. of non-zero atoms (out of 1000 concatenated atoms) with variation of No. of iterations using ASNA



(a) (b) (c)

Figure 2.8: Variation of SDR, NDR, MAE-SNR and No. of non-zero atoms (out of 1000 concatenated atoms) with variation of λ using ASNA-L1



(a) (b) (c)

Figure 2.9: Variation of SDR, NDR, MAE-SNR and No. of non-zero atoms (out of 1000 concatenated atoms) with variation of No. of iterations using ASNA-L1 and $\lambda = 1$

NDR, MAE-SNR and No. of non-zero weights for all the combinations of source recovery algorithms (ASNA, ASNA-L1, sup. NMF); and dictionary learning methods (K-medoids and SNMF). It is seen

Table 2.3: Comparison of SDR, NDR, MAE-SNR and No. of non-zero weights using K-medoid and SNMF dictionaries with weights estimated using ASNA, ASNA-L1 and sup-NMF algorithms

Recovery type	ASNA		ASNA-L1		sup.NMF	
Dictionary	K-medoid	SNMF	K-medoid	SNMF	K-medoid	SNMF
SDR (dB)	6.75	4.46	8.25	7.30	7.85	8.11
NDR (dB)	6.74	4.45	8.25	7.29	7.84	8.10
MAE-SNR (dB)	3.07	7.06	0.73	1.77	1.11	0.91
No. of non-zero weights	46.15	28.5	40.45	18.57	1000	1000

that the No. of non-zero weights in the case of sup.NMF as source recovery is non-sparse as it uses multiplicative update and L_1 regularization.

It is observed that we get the best separation performance of SDR=8.25 dB using K-medoid dictionary with ASNA-L1 and $\lambda = 1$ while the estimated SNR is closest to the original SNR of 0 dB in the case of K-medoid dictionary with ASNA-L1 recovery.

It is seen that the No. of non-zero weights in the case of sup.NMF recovery is 1000 i.e. none of the weights are zero. The reason behind this is that the weights are updated using multiplicative update. It is observed that most of the weights have very low values due to L_1 regularization which encourages sparsity. Figure 2.10 shows the plot of the histogram of the weights estimated using SNMF dictionary and sup.NMF source recovery for a mixed feature $\mathbf{y}$. It is seen that out of 1000 weights, 368 of them have value less than 0.001, 724 less than 0.01 and only 9 weights have value greater than 0.1.

In the following sections , we use the source recovery algorithms with the value of the parameters that gave the best results in this section.

2.6 Classification of noise

Given a test noise signal $y_{ns}[n]$, we need to identify the signal as belonging to one of the noise sources. Figure 2.11 shows the plot of 500 ms segments of babble, factory1, machinegun and white noise. It is seen that different noises have different signal characteristics; for example, machinegun noise has silence regions followed by impulsive bursts. We use the 15 noise dictionaries learnt as explained in Sec.2.4 and the test audio signal is classified as that source which gives the highest value for an appropriately defined objective measure. Figure 2.12 shows the plot of first three atoms learnt using TDCS-0.9 from white and babble noise sources, respectively.

2.6.1 Evaluation metrics for noise classification

The learnt dictionaries are used to extract measures for identifying a source. Given an unknown noise signal, the mag.STFT features are extracted as given in Sec.2.4.1. Since we know the dictionaries for all the sources, we estimate the following three measures for classification:

1. *Signal to distortion ratio* (SDR) [93] between the mag.STFT feature $\mathbf{y}_{ns}$ and the estimated feature, $\hat{\mathbf{y}}_{ns}^k = \mathbf{D}_{ns}^k \mathbf{x}_{ns}^k$, $1 \leq k \leq M_{ns}$ for the M_{ns} noise dictionaries. The weights for each of the noise dictionary, $\mathbf{x}_{ns}^k$ are recovered using the source recovery algorithms discussed in Sec.2.3.

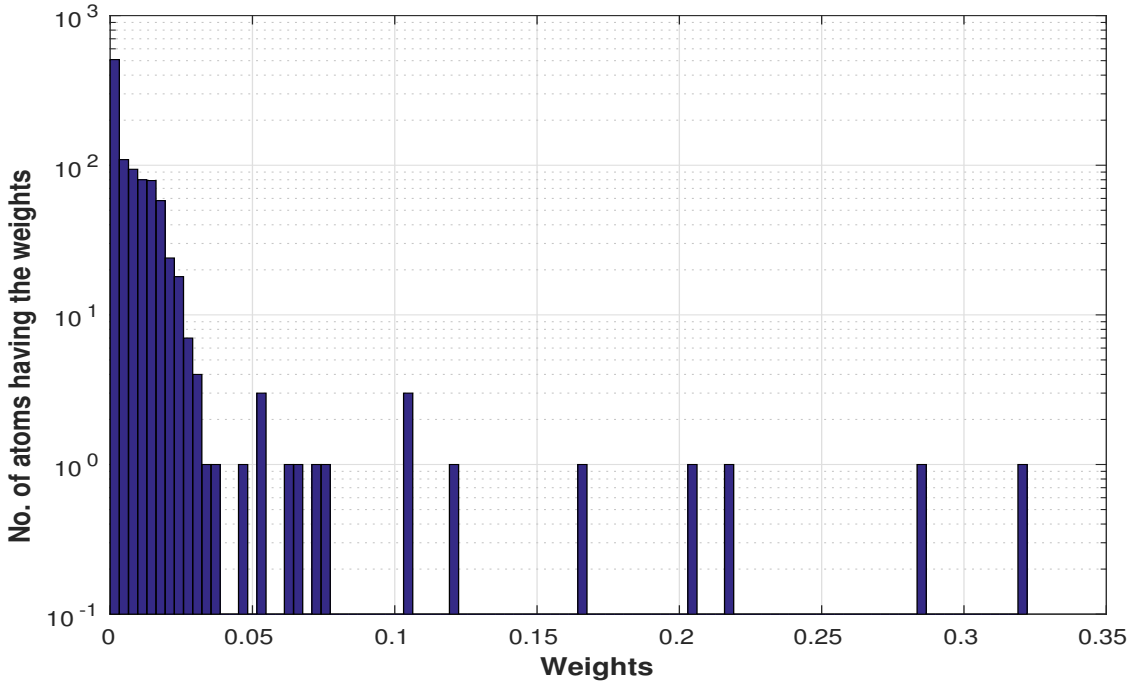


Figure 2.10: Histogram of the weights estimated using sup.NMF

The SDR with respect to the k^{th} dictionary $\mathbf{D}_{ns}^k$ is defined as :

$$SDR^k = 20 \times log_{10}(||\mathbf{y}_{ns}||_2 / ||\mathbf{y}_{ns} - \hat{\mathbf{y}}_{ns}^k||_2) \tag{2.14}$$

A feature $\mathbf{y}_{ns}^m$ belonging to the m^{th} source can be approximated to a good accuracy by atoms belonging to $\mathbf{D}^m$, since $\mathbf{D}^m$ has been learnt from the same source. So, $||\mathbf{y}_{ns}^m - \hat{\mathbf{y}}_{ns}^k||_2$ is expected to be minimum for the m^{th} source, since $\mathbf{y}_{ns}^m$ may not be approximated well by atoms from the dictionaries of other sources. Thus, the SDR^k is expected to be maximum for the m^{th} dictionary. The estimated source index $\hat{m}$ for the feature vector of each frame of the test signal is given as $\hat{m} = \arg\max_{\mathrm{k}} \ SDR^k$.

Figure 2.13 shows the plot of weights recovered using babble and white noise dictionary and the corresponding SDR for a test babble noise signal. It is seen that the SDR using babble noise dictionary is 7.55 dB as compared to 1.02 dB using white noise dictionary. It shows that we get a clearly better representation of test babble noise using babble noise dictionary, and hence SDR can be used as a metric for classification.

2. *Number of non-zero weights* (NNZ): We propose this new feature for each source in the weight vector $\mathbf{x}$ recovered using a dictionary $\mathbf{D}_{ns}$, obtained by concatenating the dictionaries of all the M_{ns} individual noise sources: $\mathbf{D}_{ns} = [\mathbf{D}_{ns}^1 \ \ \mathbf{D}_{ns}^2 ... \mathbf{D}_{ns}^{M_{ns}}]$. The vector $\mathbf{x}_{ns} = [\mathbf{x}_{ns}^{1\intercal} \ \mathbf{x}_{ns}^{2\intercal} ... \mathbf{x}_{ns}^{M_{ns}\intercal}]^\intercal$

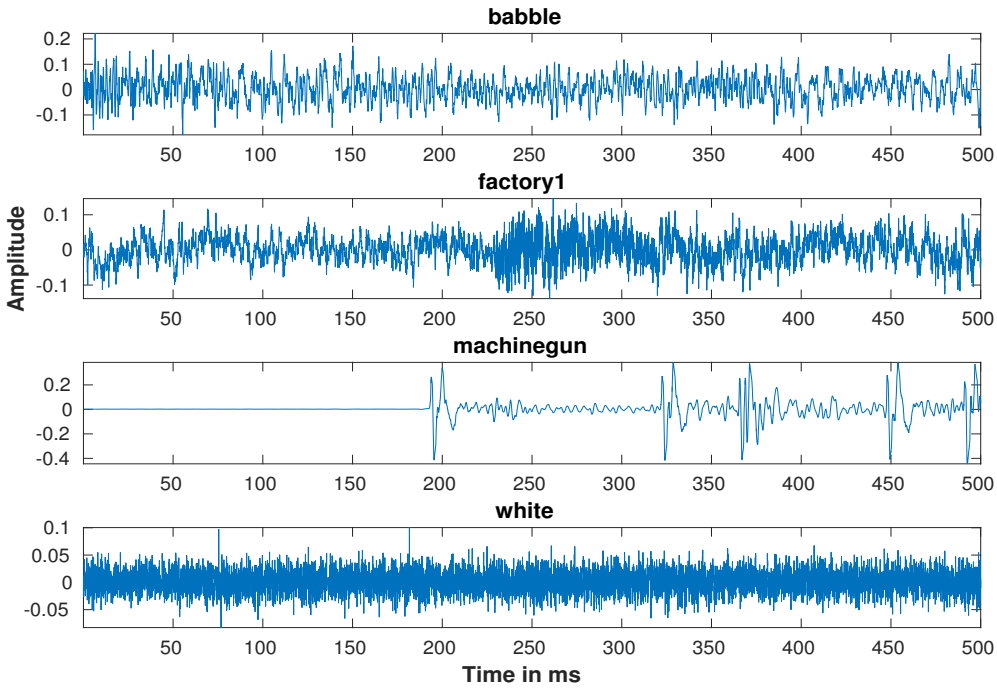
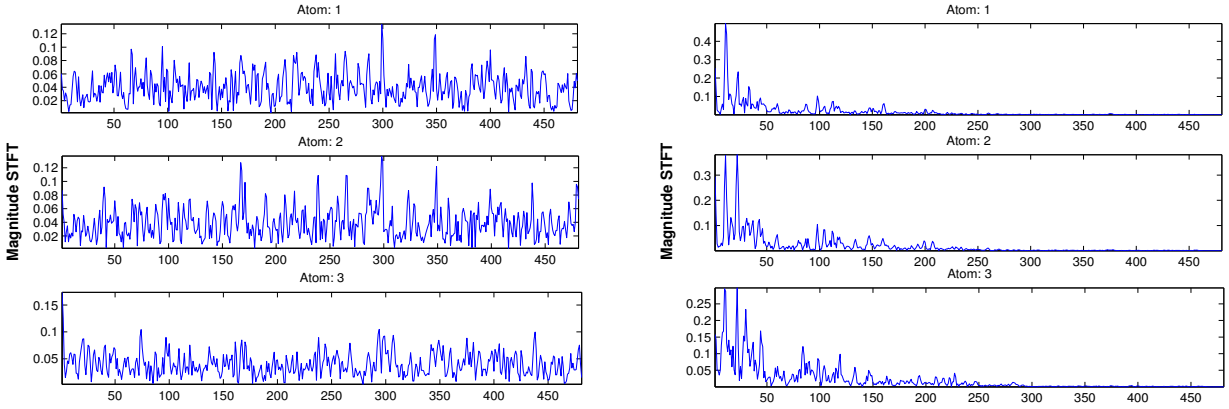


Figure 2.11: Plot of 500 ms segments of different noise signals



(a) White (b) Babble

Figure 2.12: The first three atoms from dictionaries of two noise sources learnt using TDCS-0.9

obtained by using the source recovery algorithms discussed in Sec.2.3 is a concatenation of individual weight vectors $\mathbf{x}_{ns}^k$ of M_{ns} sources, and is expected to be sparse. NNZ^k is defined as the No. of non-zero weights in the weight vector $\mathbf{x}_{ns}^k$ corresponding to the k^{th} noise dictionary.

A test feature vector $\mathbf{y}_{ns}^m$ belonging to the m^{th} source can be represented better by atoms from the m^{th} dictionary than by atoms from other dictionaries. Since $\mathbf{D}_{ns}$ contains atoms from all

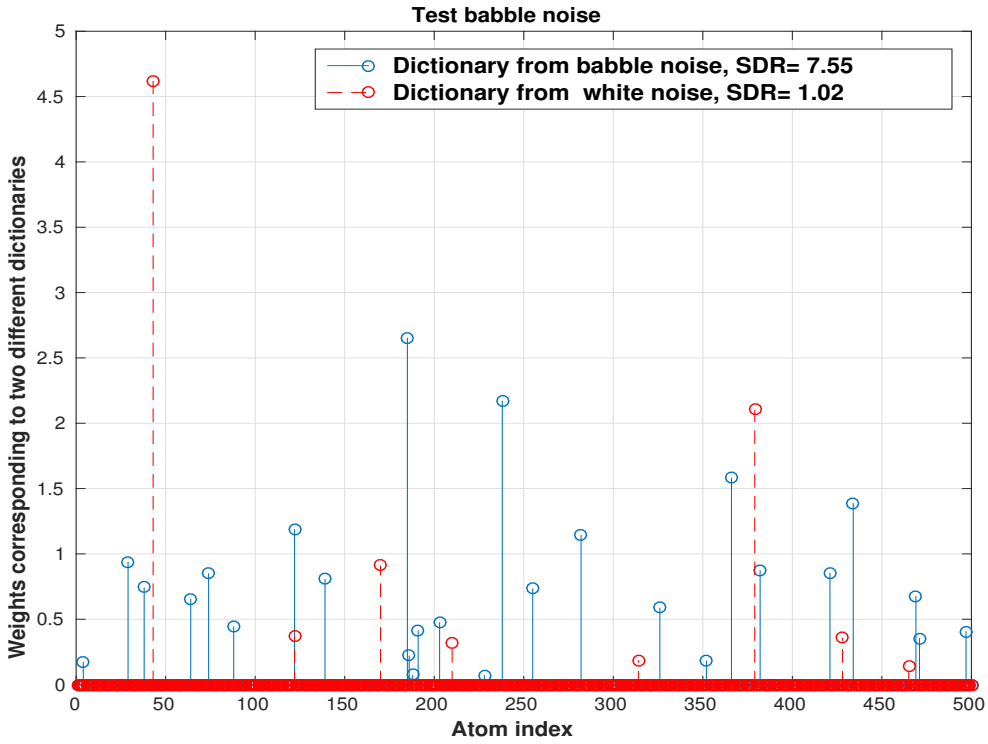


Figure 2.13: Plot of weights estimated using ASNA and SDR corresponding to babble and white noise dictionaries for a test babble noise feature

the sources, the number of non-zero weights, NNZ^m corresponding to the correct dictionary $\mathbf{D}_{ns}^m$, which is now a sub-matrix of $\mathbf{D}_{ns}$, may be expected to be higher than NNZ^k, $k \neq m$. The estimated source index $\hat{m}$ for the test vector $\mathbf{y}_{ns}^m$ is given by $\hat{m} = \arg\max_{\mathrm{k}} \ NNZ^k,\ 1 \leq k \leq M_{ns}$. The weight vector $\mathbf{x}_{ns}$ estimated using ASNA-L1 and K-medoid dictionaries is shown in Fig. 2.14(a) and is sparse for the dictionary $\mathbf{D}_{ns}$. The number of non-zero weights for each source dictionary is illustrated in Fig. 2.14(b). For a test frame of babble noise, the highest NNZ is 12 corresponding to babble noise dictionary (atom indices 1 to 500 in $\mathbf{D}_{ns}$), while 5 is the next highest for the destroyerengine dictionary. Thus, a margin of 7 or a factor of 2.4, is obtained for correct classification.

3. *Sum of weights* (SW) is another scalar measure proposed, defined as the sum of the elements of the vector $\mathbf{x}_{ns}^k$, recovered using the same concatenated dictionary, $\mathbf{D}_{ns}$. In case the weights are non-sparse, it is observed that SW^k is more reliable than NNZ^k. Figure 2.14(b) also illustrates the distribution of SW for each of the dictionaries. $\hat{m} = \arg\max_{\mathrm{k}} \ SW^k$ gives the estimated source index for a test vector $\mathbf{y}_{ns}^m$. The sum of weights is the highest (10.21) for babble noise dictionary, while the next highest is 3.06 for destroyerops, a factor of about 3.33 for correct classification.

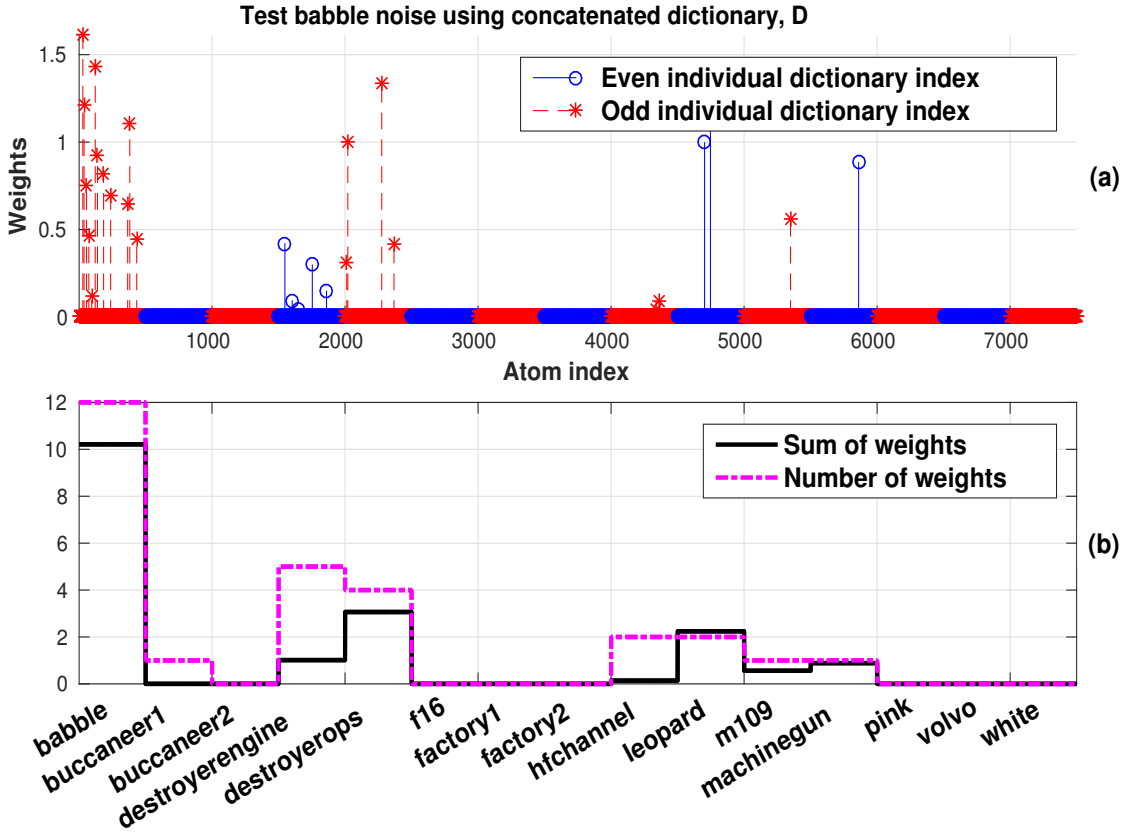


Figure 2.14: (a) Weights estimated by ASNA-L1 using concatenated dictionary, $\mathbf{D}_{ns}$ for a single test frame of babble noise . (b) Sum and number of non-zero weights in (a) for each of the K-medoid dictionaries.

It is to be noted that the dictionary used for both NNZ and SW is a concatenated dictionary $\mathbf{D}_{ns}$, while the measure SDR is derived using separate dictionaries $\mathbf{D}_{ns}^{k}$.

2.6.2 Results on noise classification

The noise classification accuracy has been evaluated on the validation set of all the fifteen noises discussed in Sec.2.4. Frames with very low energy, which may correspond to silence have been ignored for evaluating the accuracy. This is similar to what was done during dictionary learning in Sec.2.4.1. Table 2.4 shows the frame-wise classification accuracy using SDR measure on the validation set for the Variable TDCS-0.9,0.8, random selection of features, TDCS-0.9,0.8, K-medoid and SNMF dictionary learning algorithms and weights recovered using ASNA, ASNA-L1 and sup.NMF. We have also evaluated the frame-wise accuracy using a multiclass SVM classifier. We have learnt the one-vs-one multiclass SVM classifier model using the K-medoids dictionaries for each noise as training features and get an accuracy of 78.26%. It is seen that using a threshold of 0.9 (TDCS-0.9) gives better accuracy than using 0.8 (TDCS-0.8). For the SNMF dictionary, estimating the weights using the sup.NMF as in [80], results in a frame-wise accuracy of 71.06% as compared to 84.77% using ASNA algorithm. Further, the accuracy is poor when learning variable number of dictionary atoms as

in Variable TDCS-0.8,0.9. We get the highest accuracy of 97.82% using K-medoid dictionaries with ASNA.

Table 2.4: Frame-wise noise classification accuracy using SDR measure

Accuracy (%)	Variable TDCS-0.9	Variable TDCS-0.8	Random	TDCS-0.9	TDCS-0.8	K-medoid	SNMF
ASNA	94.25	82.48	97.66	97.10	95.51	97.82	84.77
ASNA-L1	93.96	83.12	97.45	96.25	94.32	97.38	82.26
sup.NMF	86.41	78.17	85.01	83.18	78.26	87.80	71.06

Table 2.5 shows the frame-wise classification accuracy using NNZ and SW measures. The accuracies using NNZ and SW are generally poorer than those using SDR measure and we get the highest accuracy of 97.91% using SW measure with sup.NMF and K-medoid dictionary. It is seen that sup.NMF recovery gives poor accuracy using NNZ for every dictionary type since all the weights are non-zero. It is observed that using SDR measure gives higher accuracy with ASNA than that of ASNA-L1 while NNZ and SW give higher accuracy using ASNA-L1. The reason is that NNZ and SW measures are obtained using a concatenated dictionary and separability is better using ASNA-L1 than ASNA as seen is Sec. 2.5.2.

Table 2.5: Frame-wise noise classification accuracy using NNZ and SW measures from concatenated dictionary

Accuracy (%)		Random	TDCS-0.9	TDCS-0.8	K-medoid	SNMF
NNZ	ASNA	79.06	76.72	73.11	77.07	85.69
	ASNA-L1	87.04	86.51	84.34	85.56	95.25
	sup.NMF	6.67	6.67	6.67	6.67	6.67
SW	ASNA	82.49	80.58	78.37	80.88	80.37
	ASNA-L1	96.59	95.84	92.06	96.54	97.43
	sup.NMF	97.47	97.55	96.45	97.91	93.92

It is seen SDR measure with ASNA recovery gives slightly lower (0.09%) noise classification accuracy than SW with sup.NMF recovery and K-medoid dictionary. As we have experimented extensively using ASNA and used sup.NMF recovery for comparison, we show further results for noise classification using ASNA recovery with SDR measure instead of sup.NMF recovery.

Table 2.6 shows the confusion matrix for noise classification accuracy as percentage of noise features classified as each of the fifteen noise classes using SDR measure with K-medoid dictionaries and ASNA recovery.

In the above discussion, we have given frame-wise accuracy. Accuracy can also be computed at the level of a cluster of contiguous frames. Two higher level measures are defined for the i^{th} dictionary, namely, accumulated SDR (ASDR) and moving ASDR (MASDR) as:

$$ASDR^i(q) = \sum_{j=1}^{q} SDR^i(j) \tag{2.15}$$

$$MASDR^i(q) = \sum_{j=q-p+1}^{q} SDR^i(j) \tag{2.16}$$

Table 2.6: Confusion matrix for noise classification accuracy in % using SDR measure with K-medoid dictionaries and ASNA. buc1: buccaneer1, buc2: buccaneer1, des.engine: destroyerengine, des.ops: destroyerops, mach.gun: machinegun

Original/Estimated	babble	buc1	buc2	des.engine	des.ops	f16	factory1	factory2	hfchannel	leopard	m109	mach.gun	pink	volvo	white
babble	99.39	0.00	0.00	0.00	0.26	0.06	0.06	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
buccaneer1	0.00	99.46	0.00	0.00	0.00	0.03	0.19	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00
buccaneer2	0.00	0.00	99.90	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00
destroyerengine	0.03	0.00	0.00	99.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
destroyerops	0.06	0.00	0.00	0.00	99.30	0.00	0.32	0.10	0.00	0.00	0.22	0.00	0.00	0.00	0.00
f16	0.00	0.00	0.00	0.00	0.00	99.23	0.03	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00
factory1	4.09	0.06	0.26	0.03	0.13	0.10	84.55	9.16	0.00	0.00	0.22	0.00	1.40	0.00	0.00
factory2	0.67	0.00	0.00	0.00	0.03	0.06	1.79	96.01	0.00	0.03	1.24	0.00	0.00	0.16	0.00
hfchannel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
leopard	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.43	0.00	0.57	0.00	0.00	0.00
m109	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.97	0.00	0.00	0.00	0.00
machinegun	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	99.94	0.00	0.00	0.00
pink	0.00	0.03	0.00	0.00	0.03	0.00	9.48	0.00	0.00	0.00	0.00	0.00	90.46	0.00	0.00
volvo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	99.87	0.00
white	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	99.97

where $ASDR^i$ is the ASDR corresponding to i^{th} noise dictionary, q is the index of the present frame and p is the number of frames accumulated.

Figure 2.15 shows the frame-wise SDR and the corresponding ASDR estimated using K-medoid dictionaries and ASNA for five test frames of factory1 noise (worst performing audio source in Table 2.6). Only two other audio sources having highest SDR's are shown, for clarity. It is seen in Fig.2.15 that even though frame-wise SDR for the second and fourth frame is lower for factory1 noise, the corresponding ASDR is higher and gives correct classification. Using MASDR, we find that 100% classification accuracy can be obtained with $p = 7$ for eleven of the noise sources implying that any set of seven consecutive frames (150 ms) of the test noise are sufficient for correct classification. Test buccaneer1 noise requires $p = 10$, and pink noise, $p = 12$ for 100% classification.

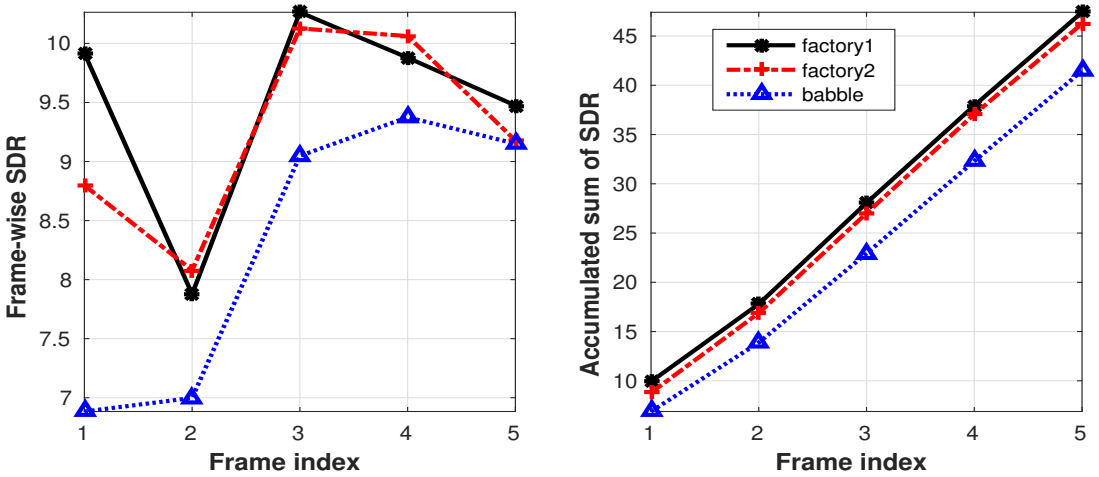


Figure 2.15: Advantage of accumulated SDR over frame-wise SDR for five frames of factor1y noise.

In a real life scenario, the accuracy of classification based on accumulated classification measures

is more relevant than individual frame level accuracy, since the classification algorithm gets a stream of test noise signal as input. So, even though a few frames may be individually misclassified, the accumulated classification measure correctly classifies the noise source.

In addition to showing classification on already known audio source classes, we have shown classification of new noise samples, each of a duration of 20 seconds recorded by us in Table 2.7. Similar results on classification of new noises have been shown in Maleh et al. [58]. We have recorded new noise samples in different background environments like bus, mess, railway station, market and metro. Given a new noise sample, we mapped each of the frames to one of the already learnt fifteen noise sources using SDR measure estimated by ASNA and K-medoid dictionaries. In the table, we have only shown seven of the already learnt noises as mapping to other noise sources is negligible. For example, recorded mess noise is classified as babble ($\approx$ 87%) and factory1 ($\approx$8%) mostly, which is reasonable since many people are speaking in a mess environment at the same time. This is very useful in the cases where we encounter a new background environment and we need to estimate its composition with respect to already learnt (known) audio classes.

Table 2.7: Distribution of frames (in %) of newly recorded noises classified as seven already learnt noise sources i.e. babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, factory1 and hfchannel using SDR measure.

Recorded noise	babble	buccaneer1	buccaneer2	destroyerengine	destroyerops	factory1	hfchannel
bus	51.58	13.16	0.23	2.18	17.82	10.98	0.00
mess	86.99	1.35	0.38	0.23	1.65	8.12	0.53
railwaystn	66.47	9.85	0.00	4.14	1.73	9.77	3.91
market	16.39	38.95	20.45	3.83	0.68	15.79	1.35
metro	16.99	54.14	11.05	0.83	0.00	0.98	14.14
mall	76.17	4.06	0.00	5.56	0.00	1.73	12.41
traffic	3.16	65.71	18.42	0.68	0.38	6.32	2.86
construction	86.54	3.68	0.00	0.00	0.15	9.55	0.08

2.6.3 Comparison with previous work

Maleh et al. [58] performed frame-wise noise identification (frame size of 20 ms) using line spectral frequencies as features. They trained the classifier using 18.75 minutes of audio data each from 5 noise classes (three of them from NOISEX database), and tested on 500 frames of data for each class. Chu et al. [60] obtained an overall accuracy of 83.9% in recognizing 14 environmental sounds. We have used 15 noise classes, and obtained an overall frame level accuracy of 97.82% using SDR and ASNA source recovery, compared to 89% reported in [58]. The highest accuracy given by majority vote classifier in [69] is around 78%.

The advantage of our approach is that we have learnt the dictionary atoms from the training data itself and classified the test signal based on how well the features are represented by weighted combination of the dictionary atoms. Other approaches have used reduced features like zero crossing rate and root-mean-square energy [58] and fixed dictionaries atoms from Gabor, Haar and Fourier dictionaries [60] which may not capture all the feature variations in the test signal.

2.7 Classification of speaker

In this section, we deal with a simple classification problem. Given a clean speech utterance, we need to identify the signal as belonging to one of the speaker sources. We use the 30 speaker dictionaries learnt in Sec.2.4 and the test utterance is classified as that source which gives the highest value for an appropriately defined objective measure.

The speaker classification accuracy has been evaluated on the validation set of all the thirty speakers discussed in Sec.2.4. Frames having very low energy have been ignored for evaluating the accuracy. The features are extracted as given in Sec.2.4.1. We evaluate speaker classification using metrics similar to the case of noise classification. We estimated weights using ASNA, ASNA-L1 and sup.NMF and used SDR and SW metrics defined in Sec.2.6.1 for speaker classification.

2.7.1 Results on speaker classification

Table 2.8 shows the speaker classification accuracy using SDR measure computed from various dictionary learning methods and weights estimated using ASNA, ASNA-L1 and sup.NMF. Here, we compute utterance level accuracy based on the maximum occurrence of frames classified as one of the speaker classes. We learn the one-vs-one multiclass SVM classifier model using the K-medoids dictionaries for each speaker source as training features and get an utterance level and frame-wise accuracy of 46.66 and 17.92%, respectively. Due to singular or badly scaled SNMF dictionaries, we did not get results for SNMF using ASNA-L1. We get a speaker classification accuracy of 86.67% (utterance level) and 24.87% (frame-wise) using SNMF dictionary and sup.NMF recovery.

Table 2.8: Utterance level (frame-wise) speaker classification accuracy using SDR.

Accuracy (%)	Variable TDCS-0.9	Variable TDCS-0.8	Random	TDCS-0.9	TDCS-0.8	K-medoid	SNMF
ASNA	36.66 (18.53)	10 (10.76)	60 (21.95)	40.00 (19.41)	66.67 (20.98)	60.00 (22.97)	56.67 (20.97)
ASNA-L1	40.00 (19.83)	13.33 (11.87)	73.33 (24.25)	43.33 (20.74)	73.33 (22.96)	80 (25.25)	-
sup.NMF	63.33 (24.30)	13.33 (15.15)	100 (27.77)	83.33 (24.46)	96.67 (26.78)	100 (28.32)	86.67 (24.87)

Table 2.9 shows the utterance level and frame-wise accuracy using SW measure estimated using the concatenated dictionary. The utterance level classification is done using sum of SW measure over all frames corresponding to each class. It is seen that we get very good accuracy using ASNA-L1. Due to singular or badly scaled SNMF dictionaries, we did not get results for SNMF using ASNA-L1.

Table 2.9: Utterance level (frame-wise) speaker classification accuracy using SW measure.

Accuracy (%)	TDCS-0.9	TDCS-0.8	Random	K-medoid	SNMF
ASNA	53.33 (20.47)	86.67 (21.26)	76.67 (24.16)	80.00 (23.85)	73.33 (23.74)
ASNA-L1	76.67 (27.59)	96.67 (29.07)	96.67 (32.42)	96.67 (32.23)	-
sup.NMF	66.67 (29.52)	76.67 (30.90)	86.67 (31.61)	80.00 (32.04)	90.00 (34.69)

As speech signals have varying energy with time, and voiced segments have higher energy than the unvoiced segments, we evaluate the accuracy using a subset of high energy frames and atoms in each speaker dictionary. It is observed that speaker classification is better in voiced frames due to the

presence of harmonics and formant structure. So, it is intuitive to vary the % of high energy frames selected and % of atoms corresponding to high energy features before normalization for evaluation of speaker classification accuracy. Table 2.10 shows the utterance level accuracy using SW measure by selecting combinations of all frames, 20%, 40%, 60% and 80% of frames having highest energy, with all atoms, 60% and 80% of atoms having highest energy for K-medoid and TDCS-0.8 dictionaries and weights estimated using ASNA-L1.

Table 2.10: Speaker classification accuracy as a function of percentage of high energy frames and high energy atoms selected using SW measure on K-medoid and TDCS-0.8 dictionaries with ASNA-L1

Accuracy (%)	All frames		20%		40%		60%		80%	
	K-medoid	TDCS-0.8	K-medoid	TDCS-0.8	K-medoid	TDCS-0.8	K-medoid	TDCS-0.8	K-medoid	TDCS-0.8
All atoms	96.67	96.67	83.33	83.33	86.67	90	96.67	93.33	96.67	96.67
60%	90	90	83.33	90	90	86.67	90	86.67	90	90
80%	96.67	93.33	83.33	83.33	86.67	93.33	100	93.33	96.67	93.33

It is seen that noise classification accuracy using TDCS is better at a threshold of 0.9 while speaker classification is better at 0.8. Thus, selection of appropriate threshold is necessary.

2.8 Speaker and noise classification/separation in noisy speech

In this section, we address the problem of both speaker and noise classification in a noisy speech signal and subsequent separation. Noisy speech signal, $y[n]$ is simulated as a linear combination of two sources, speech from i^{th} speaker source, $y^i_{sp}[n]$ and noise from j^{th} noise source, $y^j_{ns}[n]$ as in Sec.2.5:

$$y[n] = y^i_{sp}[n] + y^j_{ns}[n] \tag{2.17}$$

Features are extracted from the frames of the noisy speech signal as explained in Sec.2.4.1. The speech and noise are constrained to belong to a specific set of speakers and noise sources, and the test signal is classified as belonging to one of the predefined speaker and noise sources. Figure 2.16 shows a part of an utterance from a female speaker, babble noise and the noisy speech signal at an SNR of 0 dB . Given the noisy speech signal, we estimate the noise and the speaker source, and use the dictionaries of the estimated noise and speaker sources for source separation.

2.8.1 Classification using block sparsity and source recovery of the mixed signal

Simultaneous estimation of both speaker and noise index is difficult since there are $M_{sp} \times M_{ns}$ combinations of speech and noise sources, which can form the noisy signal $\mathbf{y}$ as:

$$\mathbf{y} \approx \hat{\mathbf{y}} = \mathbf{Dx} = [\mathbf{D}^1_{sp} ... \mathbf{D}^i_{sp} ... \mathbf{D}^{M_{sp}}_{sp} \mathbf{D}^1_{ns} ... \mathbf{D}^j_{ns} ... \mathbf{D}^{M_{ns}}_{ns}][\mathbf{x}^{1\intercal}_{sp} ... \mathbf{x}^{i\intercal}_{sp} ... \mathbf{x}^{M^\intercal_{sp}}_{sp} \mathbf{x}^{1\intercal}_{ns} ... \mathbf{x}^{j\intercal}_{ns} ... \mathbf{x}^{M^\intercal_{ns}}_{ns}]^\intercal \tag{2.18}$$

Equation 2.18 shows the block structure representation of $\hat{\mathbf{y}}$ as linear combination of the blocks of $\mathbf{D} = [\mathbf{D}^1_{sp} ... \mathbf{D}^i_{sp} ... \mathbf{D}^{M_{sp}}_{sp} \mathbf{D}^1_{ns} ... \mathbf{D}^j_{ns} ... \mathbf{D}^{M_{ns}}_{ns}]$ weighted by $\mathbf{x} = [\mathbf{x}^{1\intercal}_{sp} ... \mathbf{x}^{i\intercal}_{sp} ... \mathbf{x}^{M^\intercal_{sp}}_{sp} \mathbf{x}^{1\intercal}_{ns} ... \mathbf{x}^{j\intercal}_{ns} ... \mathbf{x}^{M^\intercal_{ns}}_{ns}]^\intercal$.

Since the test noisy signal is assumed to have only two sources, the features extracted can be

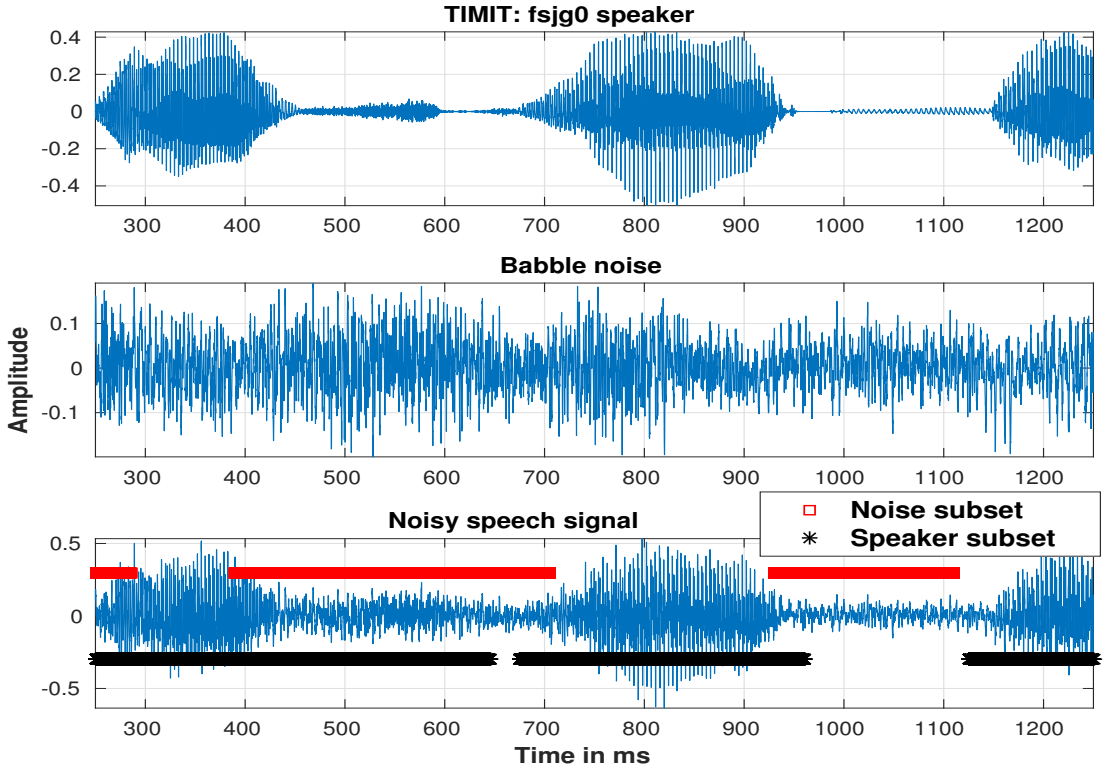


Figure 2.16: Illustration of speech, noise and the noisy speech signal.

approximated as a linear combination of atoms belonging to two source dictionaries. Assuming the i^{th} speaker and j^{th} noise source are active in $\mathbf{y}$, we use the concept of block sparsity [56] where all the blocks of $\mathbf{x}$ are zero except for $\mathbf{x}_{sp}^{i}$ for the i^{th} speaker source and $\mathbf{x}_{ns}^{j}$ for the j^{th} noise source. This reduces Equation 2.18 to:

$$\hat{\mathbf{y}} = [\mathbf{D}_{sp}^{i}\ \mathbf{D}_{ns}^{j}][\mathbf{x}_{sp}^{i\intercal}\ \mathbf{x}_{ns}^{j\intercal}]^{\intercal} \tag{2.19}$$

where $\mathbf{D}_{sp}^{i}, \mathbf{D}_{ns}^{j}$ are the dictionaries for the i^{th} speaker and j^{th} noise source; i,j and the weight vectors $\mathbf{x}_{sp}^{i}$, $\mathbf{x}_{ns}^{j}$ are unknown.

We use this concept of block sparsity for estimating the noise source index $\hat{j}$ first, and then use the concatenation of the dictionary of the estimated noise source index and all the speaker dictionaries to estimate the speaker index as $\hat{i}$.

2.8.1.1 Noise classification stage

In Sec.2.6, we have shown frame-wise classification of an audio signal containing a single noise source. We have shown how accumulated classification is useful since the noise source is the same across various frames over a significant span of time. Here, we address noise classification over the whole

noisy speech signal.

The speech component in a noisy speech signal consists of voiced, unvoiced and silence segments. So, when speech and noise sources are mixed at a particular SNR, the frames containing silence segments of speech contain noise only frames. We need to identify frames containing dominant noise, which are reliable for noise classification. It is observed that using all the frames for noise classification results in a poor accuracy of 44.9% on noisy speech signals simulated at SNR= 20 dB using all speaker and noise combinations, as compared to 96.7% using a subset of the features. It is seen that for all the noisy utterances which were misclassified (when all the frames are used), they are confused with babble noise, which seems very intuitive, as speech dominant frames are similar to babble noise. Around 61.7% of the noisy utterances are classified as babble noise for all the noise types.

Figure 2.16 shows the illustration of the segments in a noisy speech signal (noise subset) which are selected for noise classification at an SNR of 0 dB. It is seen that 159 out of 287 frames i.e. 55.4% of total frames have been selected for noise classification.

Each test feature is explored for maximum cosine similarity with an atom of one of the noise and speaker dictionaries. The cosine similarity is computed with each atom of each of the noise and speaker dictionaries. Then, a subset of features of the noisy speech are selected for noise classification based on the maximum similarity between the noisy feature $\mathbf{y}$ with any of the dictionary atoms from all the noise and speaker dictionaries as:

$$\hat{k} = \arg\max_{k} \ \mathbf{y}^{\intercal}\mathbf{d}_n^k \ \forall \ 1 \leq n \leq N; \ 1 \leq k \leq (M_{ns} + M_{sp}) \tag{2.20}$$

where $\mathbf{d}_n^k$ is the n^{th} atom corresponding to the k^{th} source. If the atom with the highest correlation belongs to source $\hat{k}$, then the corresponding feature $\mathbf{y}$ is selected for noise classification, if and only if $\hat{k}$ corresponds to a noise source.

For the subset of features selected for noise classification, we use the TDCS-0.9 and K-medoid dictionary with SDR measure and ASNA recovery for classification. Figure 2.17 shows the variation of percentage of frames selected for noise classification over all the combinations of noise and speaker sources as a function of input SNR. The percentage of frames selected decreases from 87% at -10 dB to 22% at 20 dB for both TDCS-0.9 and K-medoid dictionaries. Algorithm 3 list the steps for noise classification from noisy speech. After selection of a subset of features, we estimate the noise index for each feature and find the sum of SDR^j over the frames whose noise index is estimated as j. This total SDR ($TSDR^j$) gives a higher confidence measure and the estimate of the noise index for the noisy speech is assigned as $\hat{j}$.

Algorithm 3 *Noise classification in noisy speech*

1: Given the test noisy features $\mathbf{y}$, select the subset of features which possess the highest correlation with some atom from a noise dictionary. This step is used only for selecting a subset of test features and the noise class information is abandoned.

2: A noise label is assigned to each of the subset of features using the SDR measure estimated using ASNA (Sec.2.6).

3: Find the sum of SDR^j over all the subset of features with estimated noise index j, as $TSDR^j$. Compute this for each of the noise indices.

4: The noise index for the utterance is estimated as $\hat{j} = \arg\max_j TSDR^j$

end

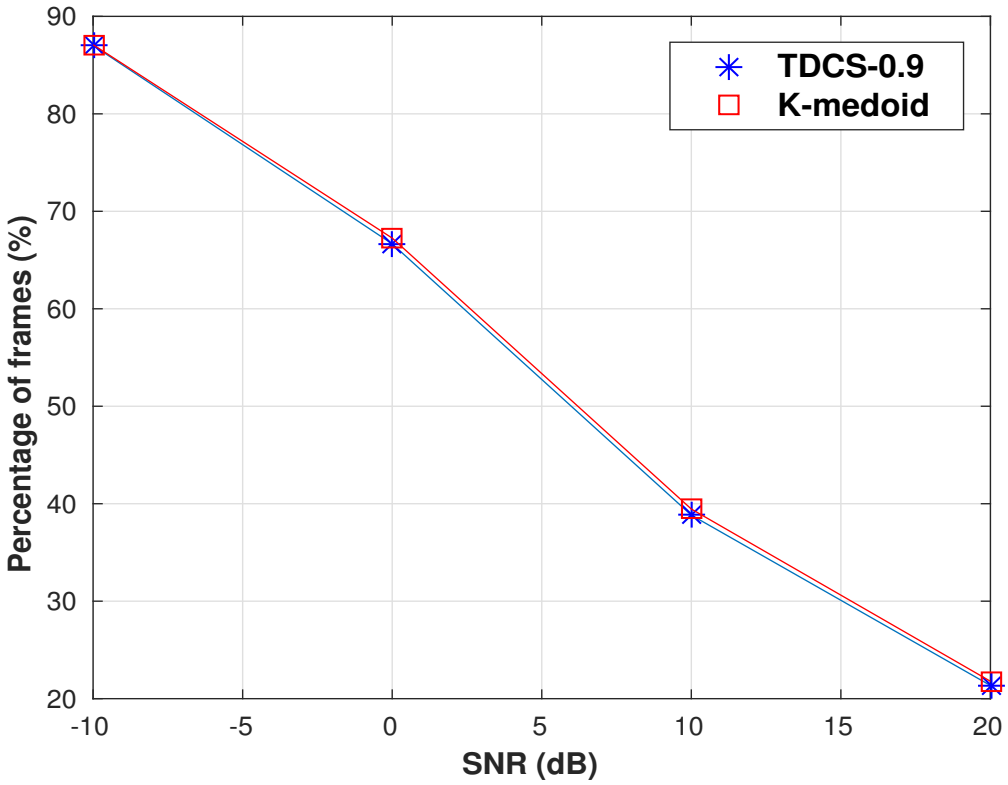


Figure 2.17: Percentage of frames selected for noise classification for various SNR's

2.8.1.2 Speaker classification stage

Given the estimated noise index, the next task is to estimate the speaker corresponding to the utterance within each noise segment. As seen in Sec.2.7, we get 100% utterance level speaker classification accuracy on the validation set using SDR measure and K-medoid dictionaries with sup.NMF recovery, and a combination of 60% of test frames and 80% of dictionary atoms having highest energy using SW measure and K-medoid with ASNA-L1. Since the speech component in noisy speech also has silence segments, those segments will be noise only segments and most likely will be low energy frames in case the noise energy does not vary much with time. So, we have selected a combination of 60% of

test frames and 80% of atoms having highest energy before normalization for speaker classification. Also, initial experiments on noisy speech [45] have suggested that using a subset of frames having high energy and using SW measure gives better classification than using all frames or SDR measure. We have seen an improvement of around 4% in speaker classification accuracy using a subset of frames and atoms over using all frames and atoms for K-medoid dictionary at SNR=0 dB. A subset of frames (60%) having high energy is shown in Fig.2.16 as segments marked as speaker subset. We compare TDCS-0.8 and K-medoids for speaker classification. It is to be noted that we use TDCS-0.9 and ASNA for noise classification whereas TDCS-0.8 and ASNA-L1 for speaker classification.

The test feature $\mathbf{y}$ is approximated as the linear combination of the dictionary atoms from the estimated noise source, $\mathbf{D}_{ns}^{\hat{j}}$ and the concatenation of speaker source dictionaries $[\mathbf{D}_{sp}^{1}....\mathbf{D}_{sp}^{M_{sp}}]$. The No. of dictionary atoms in each source dictionary is 80% of the total No. of atoms having high energy. The speaker index is determined by comparing the weights estimated in the representation:

$$\mathbf{y} \approx [\mathbf{D}_{sp}^{1}....\mathbf{D}_{sp}^{M_{sp}}\mathbf{D}_{ns}^{\hat{j}}][\mathbf{x}_{sp}^{1\intercal}... \ \mathbf{x}_{sp}^{M_{sp}^{\intercal}} \ \ \mathbf{x}_{ns}^{\hat{j}\intercal}]^{\intercal} = \mathbf{D}\mathbf{x} \tag{2.21}$$

The weight vector, $\mathbf{x}$ is estimated by minimizing the distance $dist(\mathbf{y}, \mathbf{D}\mathbf{x})$ using ASNA-L1 (Equation 2.5), where $dist()$ is the KL-divergence between $\mathbf{y}$ and $\mathbf{D}\mathbf{x}$. The measure *Sum of Weights (SW)* for each of the selected features $\mathbf{y}$ is defined as the sum of elements of $\mathbf{x}_{sp}^{k}$, $1 \leq k \leq M_{sp}$,

$$SW^{k} = ||\mathbf{x}_{sp}^{k}||_{1} \tag{2.22}$$

The steps for speaker classification are listed in Algorithm 4. SW^{k} is summed over all the frames of the subset to get an overall confidence measure, TSW^{k} to estimate the speaker index as $\hat{i}$.

Algorithm 4 *Speaker classification in noisy speech*

1: Given the estimated noise source $\hat{j}$, select a subset of features from the noisy speech as 60% of the total No. of features having highest energy.
2: Find the weights recovered using ASNA-L1 and dictionaries with high energy atoms as per Equation 2.21 for the subset of features picked up in the previous step.
3: The sum of weights, SW^{k} corresponding to each speaker dictionary index k is found as per Equation 2.22.
4: Find the total sum of weights using for each speaker source as $TSW^{k} = \sum SW^{k}$ over all the features of the subset.
5: Estimate the speaker index as $\hat{i} = \arg\max_{k} TSW^{k}$.

end

2.8.2 Speaker and noise dictionary update

A novel algorithm is proposed to update speaker and noise dictionaries. It is a generalized algorithm which works with any dictionary learning algorithm. In the case of a test signal containing noise

or speech only segments, the corresponding features are used to update the estimated noise/speaker source dictionary. The intuition behind this method is that even though the dictionary for a particular source is not a good representation, it can be considered as the base dictionary which we update using the test signal itself. It is also useful when the test signal is of short duration or it varies with time. Algorithm 5 gives the steps used for the dictionary update, where the test features $\mathbf{Y}$ are concatenated to the old dictionary $\mathbf{D}$ to update it as $\mathbf{D}_{upd}$.

Algorithm 5 *Dictionary update*

1: Given an input dictionary $\mathbf{D}$ and the test features $\mathbf{Y}$
2: Update the dictionary $\mathbf{D}$ as $\mathbf{D}_{upd} = f([\mathbf{D}\ \mathbf{Y}])$, where $f()$ is any dictionary learning algorithm like $K - medoids$

end

2.8.3 Results on speaker/noise classification and source separation performance

The noise and speaker databases used for simulating the noisy speech signals have been described in Sec.2.4. A linear combination of the test set of the noise and speaker sources simulates the noisy speech. The case where the speaker and/or noise sources are unknown is explored by considering out of set speaker/noise sources. Performance improvement is explored by adapting the speaker and noise sources using test set itself.

2.8.3.1 Testing setup

A noisy speech signal is simulated by adding test utterance from a speaker to the noise source at SNR's of -10, 0, 10 and 20 dB. As there are 15 noise and 30 speaker sources, 15×30 combinations of noisy speech signals are used for testing at different SNR's.

For testing our classification and separation performance using ground truth and updated dictionaries, we have five test cases using different combinations of dictionaries.

1. *Complete speaker and noise dictionary :* The test mixed audio signal is tested using all the noise and speaker dictionaries, and the separation is achieved using the identified noise/speaker dictionary.

2. *Ground truth speaker and noise dictionary:* The ground truth speaker and noise dictionaries are used to obtain the separation.

3. *Out of set noise sources:* The test noisy signal is tested using all the noise dictionaries except for the dictionary corresponding to the noise source used in the test signal and all the speaker dictionaries. So, by pruning ground truth noise dictionary, the test signal is tested against out of set/unknown noise sources and known speakers. The results reported using this case show the robustness of our method given unknown test noises.

4. *Out of set speaker sources:* The dictionary corresponding to the speaker source used in the test signal is removed from the training set for classification and separation. This case shows the robustness of our method given unknown test speakers.

5. *Out of set speaker and noise sources* The dictionaries corresponding to both the speaker and noise sources used in the test audio signal are removed from the training set for classification and separation. This case shows the robustness of our method for the unsupervised case of unknown speaker and noise sources.

The out of set noise and speaker dictionaries used in the test cases (3, 4, 5) give the estimated noise and speaker source indices. The dictionaries corresponding to the estimated source indices are updated using the dictionary update method given in Sec.2.8.2. For noise dictionary update, features from a 10 second segment of the validation set are used to update the estimated noise dictionary. Features from the utterance in the validation set are used to update the estimated speaker dictionary. Results on speaker classification and separation performance are reported both before and after the dictionary update using the estimate of noise/speaker index from the out of set dictionaries and updated dictionaries.

2.8.3.2 Performance results

The results for noise and speaker classification, SDR, NDR and error in estimated SNR are reported in this section. All the results given below are averages over all the combinations of speakers and noise sources.

Table 2.11 shows the overall noise classification accuracy at various SNR's of noisy speech. It is seen that we get 100% noise classification accuracy at an SNR of -10 dB and a good accuracy of 96.89% at 20 dB for both TDCS-0.9 and K-medoid dictionaries. It is observed that machinegun noise gets misclassified with some combinations of speaker sources at higher SNR's as machinegun has many silence segments, which have no discriminatory information by themselves or when corrupted with speech.

Table 2.11: Noise classification accuracy using TDCS-0.9 and K-medoid dictionary learning methods at SNR of -10, 0, 10 and 20 dB

SNR	-10 dB	0 dB	10 dB	20 dB
TDCS-0.9	100.00	99.11	97.33	96.89
K-medoid	100.00	99.78	97.78	96.89

Table 2.12 shows the overall speaker classification accuracy using the complete noise and speaker dictionaries, out of set and updated noise dictionaries. K-medoid dictionaries give better speaker classification accuracy than TDCS-0.8. Unknown noise using out of set noise dictionaries results in lowest accuracies, while using an updated dictionary entails similar accuracies as the complete dictionaries.

Table 2.12: Speaker classification accuracy using the complete noise/speaker dictionary (Complete), out of set noise dictionary (Unknown) and updated noise dictionary (Updated) at SNR values of -10, 0, 10 and 20 dB

SNR		-10 dB	0 dB	10 dB	20 dB
Complete	TDCS-0.8	34.44	65.78	82.00	88.89
	K-medoid	36.89	75.56	89.11	92.89
Unknown	TDCS-0.8	19.56	50.22	79.33	90.44
	K-medoid	18.22	57.33	87.56	92.67
Updated	TDCS-0.8	30.44	66.00	81.78	89.56
	K-medoid	33.11	74.22	88.89	92.89

Figure 2.18 shows the speaker classification accuracy in the presence of various noise classes. We get around 93% accuracy with machinegun and volvo noise using K-medoid dictionaries even at the SNR of 0 dB.

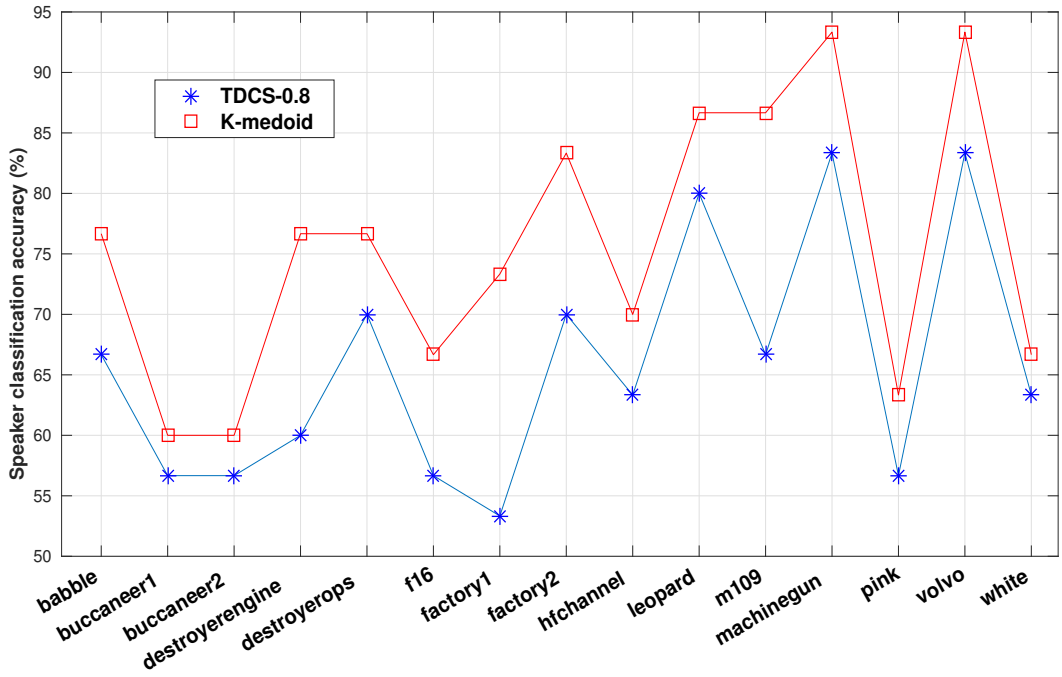


Figure 2.18: Speaker classification performance in the presence of various noise sources at SNR of 0 dB

We evaluate the source separation performance using ASNA-L1 and sup.NMF recovery methods. We use the source separation method explained in Sec.2.5 by replacing the original speaker and noise index i, j with the estimated speaker and noise index $\hat{i}, \hat{j}$ in Equation 2.6. For sup.NMF recovery, we use the SNMF dictionary for separation since it gives better separation performance as seen in Table 2.3, while the estimates of noise and speaker indices $\hat{i}, \hat{j}$ are obtained using TDCS or K-medoid. Figures 2.19 and 2.21 show the variation of SDR and NDR in dB using the combination of TDCS and K-medoid dictionary methods for classification with ASNA-L1 and sup.NMF recovery evaluated on complete dictionaries (Complete), ground truth dictionaries (Ground), out of set noise (OS noise),

out of set speaker dictionaries (OS speaker), updated noise (Upd. noise), updated speaker dictionary (Upd. speaker), out of set noise and speaker dictionary (OS noise, speaker) and updated noise and updated speaker dictionaries (Upd. speaker, noise). It is seen that using Ground gives the best SDR while using OS noise and speaker dictionaries gives the lowest SDR followed by OS noise dictionary. Upd. noise dictionary gives SDR comparable to Complete test cases. It is observed that using OS speaker does not degrade SDR much as compared to Complete at all SNR's.

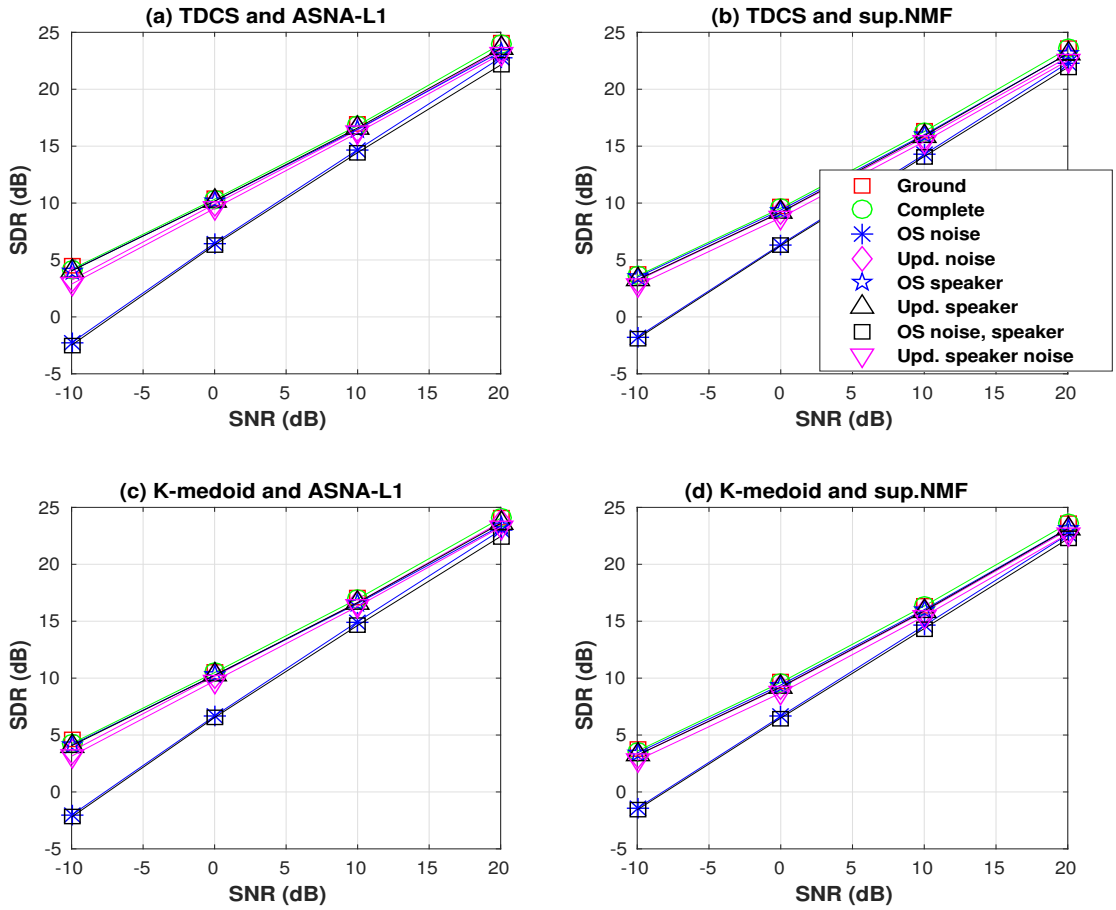


Figure 2.19: Plots of SDR as a function of input SNR using complete dictionaries (Complete), ground truth dictionaries (Ground), out of set noise (OS noise) , out of set speaker (OS speaker), updated noise dictionary (Upd. noise) and updated speaker dictionary (Upd. speaker), out of set noise and speaker dictionary (OS noise, speaker) and updated noise and updated speaker dictionary (Upd. speaker, noise) for four different combinations of dictionary and source recovery types.

SDR using K-medoid and ASNA-L1 recovery is the highest among all the four combinations of dictionary learning and source recovery methods. Table 2.13 shows the variation of SDR with input SNR. Figure 2.20 shows the same for speech mixed with each of the noise sources at 0 dB SNR.

Figure 2.22 shows the variation of mean absolute error in estimated SNR (MAE-SNR) while Fig.

Table 2.13: SDR using K-medoid and ASNA-L1 recovery method which gives the best separation performance

SDR	-10 dB	0 dB	10 dB	20 dB
Ground	4.62	10.54	17.03	24.14
Complete	4.25	10.49	16.96	24.04
OS noise	-2.03	6.68	14.96	23.00
Upd. noise	3.58	10.16	16.64	23.71
OS speaker	4.13	10.25	16.55	23.37
Upd. speaker	4.06	10.25	16.64	23.59
OS noise, speaker	-2.21	6.54	14.63	22.44
Upd. speaker, noise	3.15	9.75	16.30	23.25

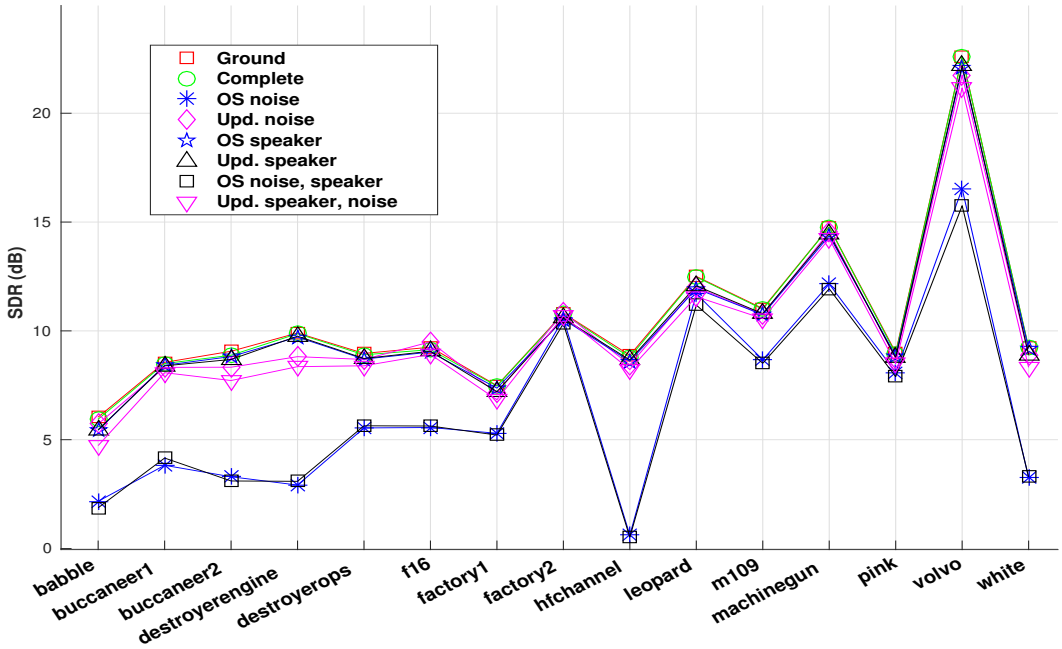


Figure 2.20: SDR for various noise sources at an SNR of 0 dB using K-medoid dictionary and ASNA-L1 recovery

2.23 shows the standard deviation of error in estimated SNR (STD-SNR) for different dictionary learning methods and test cases similar to Fig. 2.19. OS noise and OS noise,speaker results in the worst performance of all test cases, giving an MAE-SNR of around 10 dB at -10 dB. Also, we get MAE-SNR of around 0.53 dB as the best over all the test cases.

2.8.4 Comparison with previous work

The results presented are not directly comparable to other methods in the literature since the test cases simulated in our work are different. A few of the results are compared indirectly here. Joder et al. [74] reported a speaker classification accuracy of 98.9 % for eight speakers with clean speech, while we achieve 100% accuracy for 30 speakers using K-medoid dictionary on clean speech . Rose et al. [94] proposed speaker identification in noise using Gausssian mixture models for 16 speakers and

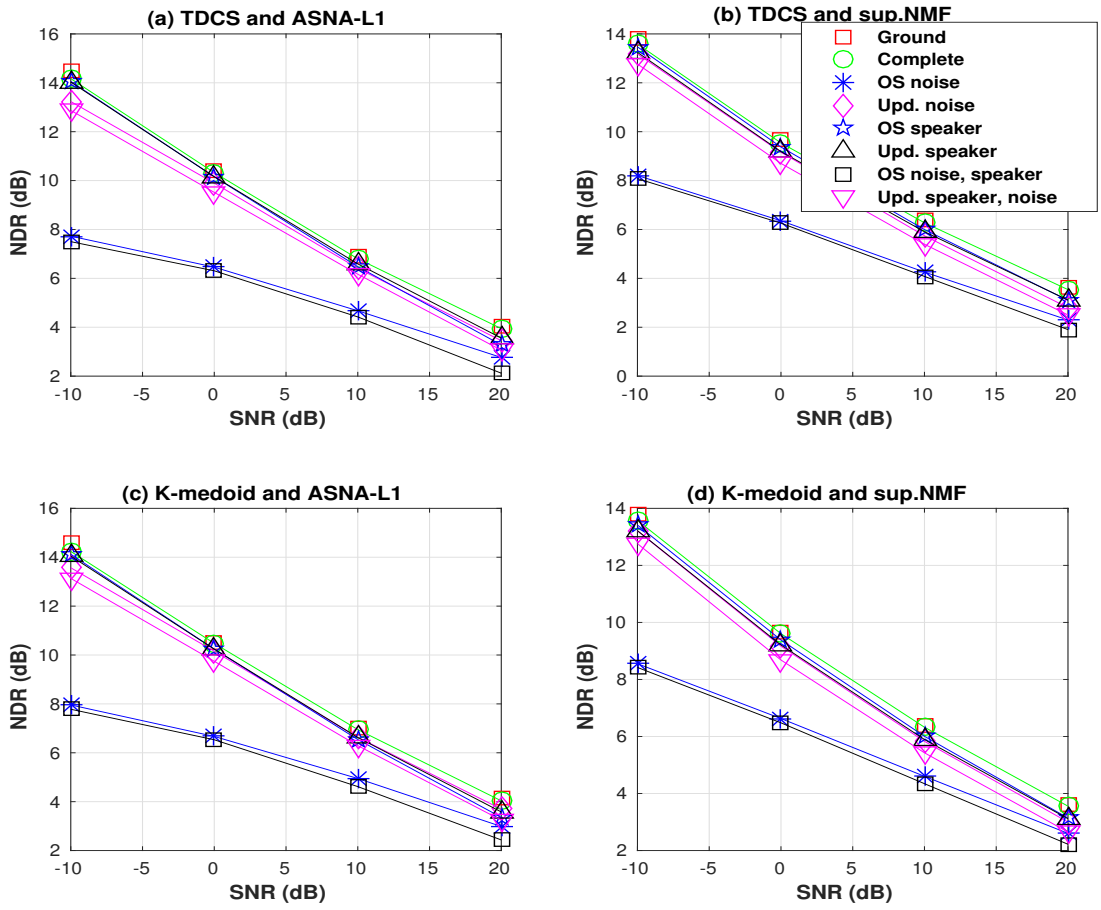


Figure 2.21: Plots of NDR for input SNR's of (a) -10, (b) 0, (c) 10 and (d) 20 dB using complete dictionaries (Complete), ground truth dictionaries (Ground), out of set noise (OS noise) , out of set speaker (OS speaker), updated noise dictionary (Upd. noise), updated speaker dictionary (Upd. speaker), out of set noise and speaker dictionary (OS noise, speaker) and updated noise and updated speaker dictionary (Upd. speaker, noise) for four different combinations of dictionary and source recovery types.

10 second segments reporting an accuracy of 79.9 % at 10 dB SNR while we get average accuracy of 93.3% tested against 30 speakers at 10 dB in the presence of babble noise. On white noise, [94] reported 68.8% accuracy, while we get 83.3% accuracy at 10 dB SNR. So, we achieve a comparably higher speaker classification accuracy. Hurmalainen et al. [95] reported speaker recognition using convolutive sparse coding on noisy speech. Loizou [4] performed speech enhancement and reported an improvement in segmental SNR in speech with speech-shaped noise of around 5 dB at 0 dB SNR while we achieve a high SDR of around 10 dB at 0 dB SNR, which is equivalent to an improvement of 10 dB. Mohammadiha et al. [6] carried out unsupervised speech enhancement based on Bayesian formulation of NMF and reported SDR of around 5.5 dB at 0 dB SNR while we achieve SDR of around 10 dB using our methods in all test cases except for OS noise and speaker case.

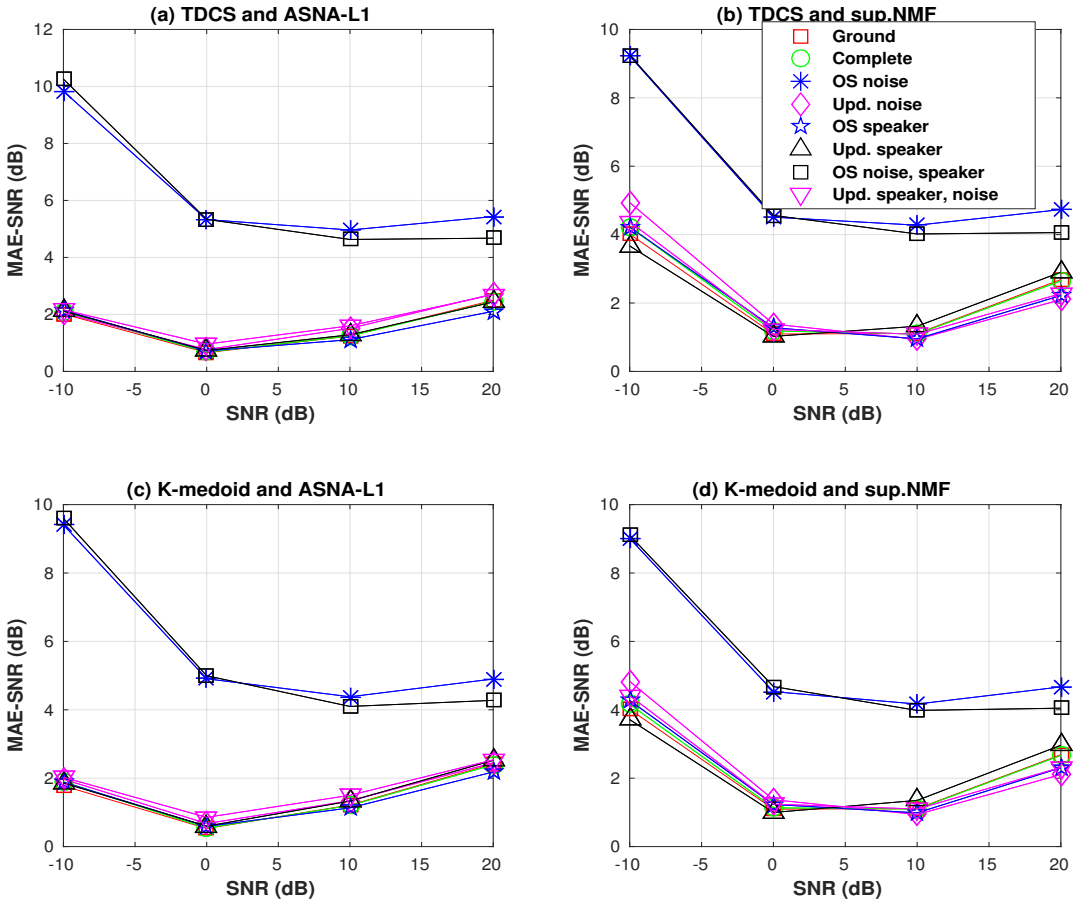


Figure 2.22: Plots of mean absolute error between the original and the estimated SNR (MAE-SNR) for input SNR of (a) -10, (b) 0 , (c) 10 and (d) 20 dB. Notations are the same as in Fig. 2.19.

2.9 Overlapped speech

In the previous section, we saw classification and separation of speech and noise in a noisy speech. In this section, we consider the case where the noise is speech from another speaker. Overlapped speech is simulated as a linear combination of utterances spoken by two speakers as:

$$y[n] = y^i_{sp}[n] + y^j_{sp}[n] \tag{2.23}$$

where $y^i_{sp}[n]$ and $y^j_{sp}[n]$ are the utterances from the i^{th} and j^{th} speakers, denoted as S1 and S2, respectively.

$y[n]$ is the overlapped speech, which can be seen as a noisy speech signal assuming the utterance from the second speaker, S2 as an interference or noise. Figure 2.24 shows a segment of overlapped speech containing a mixture of speech from a male and a female speaker.

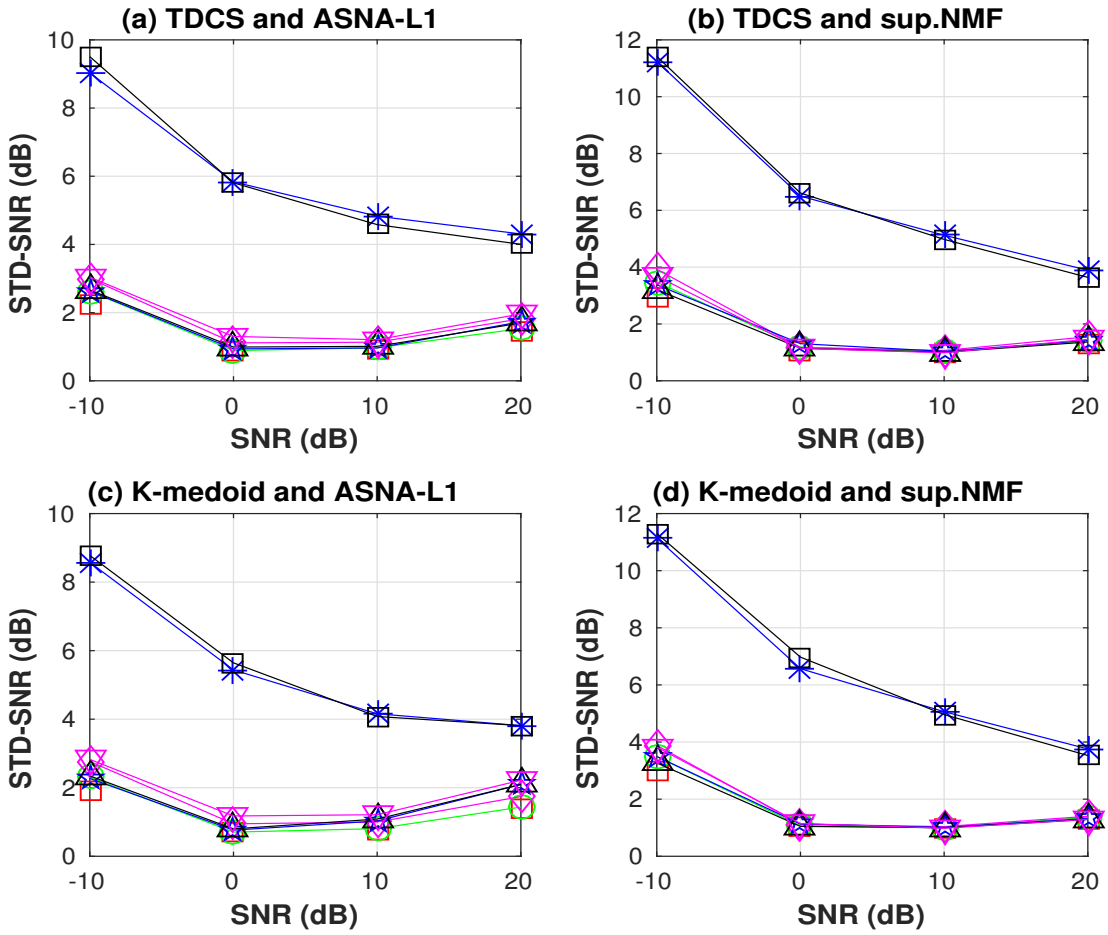


Figure 2.23: Illustration of standard deviation of error between the original and the estimated segmental SNR (STD-SNR). Notations are the same as in Fig. 2.19.

2.9.1 Classification stage

Since the overlapped speech contains two speakers, we use a concatenation of dictionaries learnt from different speakers as $\mathbf{D}_{sp} = [\mathbf{D}_{sp}^{1}...\mathbf{D}_{sp}^{i}...\mathbf{D}_{sp}^{M_{sp}}]$, where M_{sp} is the total No. of speakers. Algorithm 6 lists the steps used to classify both the speakers in any overlapped speech. Given the overlapped speech, we extract the features and classify the two speakers based on the total sum of weights over all the frames, TSW^{k}. Here, we explore a secondary level classifier where we use the dictionaries corresponding to the P highest TSW^{k} from the first classifier. Now, the speaker indices corresponding to the two highest $TSW^{k'}$ correspond to either of the ground truth indices i, j of the two speakers i.e. S1 may correspond to either the highest or the second highest $TSW^{k'}$, mostly depending upon the input S1S2R. So, we are able to classify both the speakers but the mapping of the estimated to the ground truth indices may be swapped, if the S1S2R<0 dB. In the case of more than two speakers

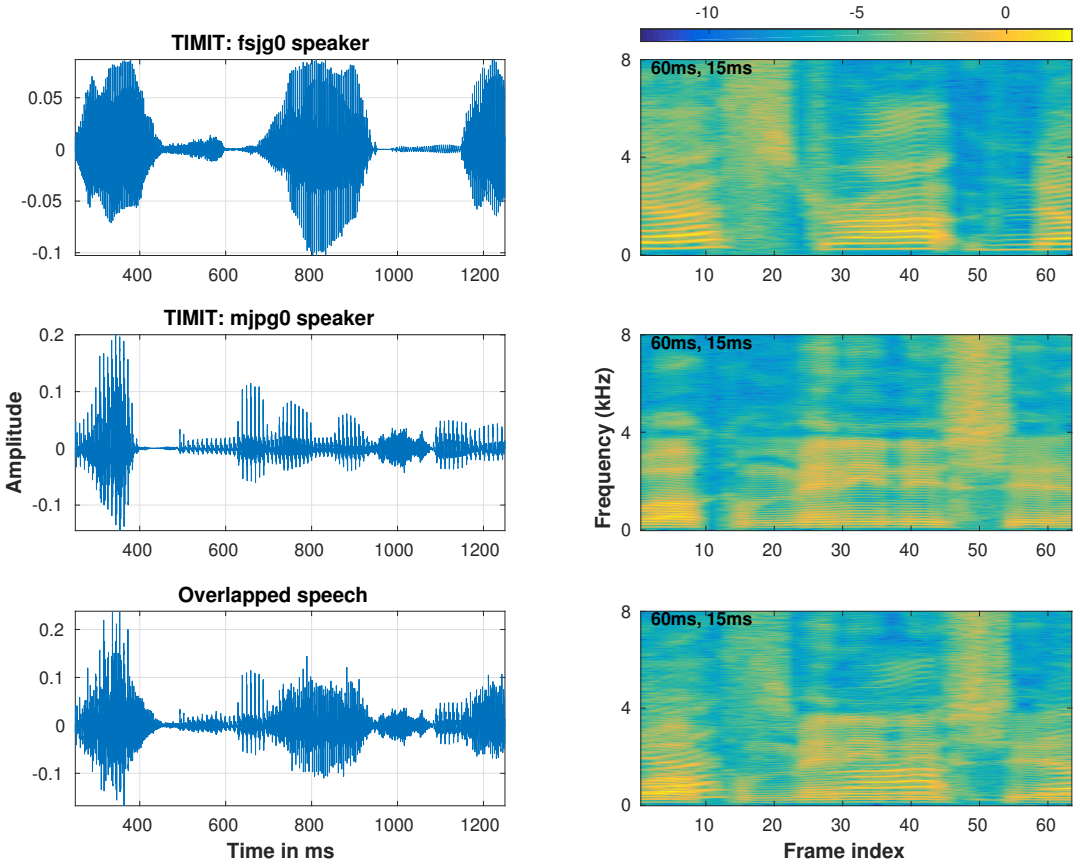


Figure 2.24: Illustration of speech segments in time domain (left) and the corresponding log spectrogram (right) for two speaker sources, TIMIT female speaker fsjg0 (S1) and TIMIT male speaker mjpg0 (S2), and the overlapped speech at a S1S2R of 0 dB.

in the overlapped speech, we can pick indices corresponding to the M highest total sum of weights in the case of a mixture of speech from M speakers.

Algorithm 6 *Speaker classification from overlapped speech*

1: Given the overlapped speech utterance $y_{sp}[n]$, extract the features $\mathbf{y}_{sp}$
2: Extract the weights corresponding to the concatenated dictionary, $\mathbf{D}_{sp}$ as $\mathbf{x}_{sp}$
3: Find sum of weights corresponding to each speaker dictionary $\mathbf{D}_{sp}^{k}$ as $SW^{k} = ||\mathbf{x}_{sp}^{k}||_1$, summed over all the frames as TSW^{k}
4: Pick the P speaker indices corresponding to the P highest TSW^{k}
5: Using the dictionary, $\mathbf{D}_{sp'}$, the concatenation of P speaker dictionaries in the previous step, find the $TSW^{k'}$ using steps 2 and 3
6: Pick the indices corresponding to the two highest $TSW^{k'}$ mapped to the original indices as the estimated speaker indices, $\hat{i}$, $\hat{j}$

end

2.9.2 Experimental setup and results

The speakers for testing the performance on overlapped speech are the same as the test set in Sec.2.4, i.e. 15 male and 15 female speakers from the TIMIT database. The overlapped speech is simulated as the following combinations (utterance from speaker S1 mixed with S2):

- One utterance each from a randomly selected female and male speakers.
- One utterance each from two male speakers randomly selected from two sets of 7 speakers each.
- One utterance each from two female speakers, selected one each from two sets of randomly selected 7 speakers each.

Thus, there are a total of 29 combinations for simulating the overlapped speech. The K-medoid and TDCS-0.8 speaker dictionaries learnt as in Sec.2.4 are used.

Since the setup is different from a noisy speech signal, we define measures similar to SNR, SDR and NDR as:

- *S1S2R*: Ratio of energy in decibel of speech from speaker 1 (S1) to speaker 2 (S2) in overlapped speech (similar to SNR)
- *S1DR*: Ratio of energy in decibel of speech to the distortion between original and estimated speech from speaker 1 (S1)
- *S2DR*: Ratio of energy in decibel of speech to the distortion between original and estimated speech from speaker 2 (S2)

We present the results on the classification of speakers S1 and S2 and the speaker source separation after recovering the estimated speaker indices $\hat{i}$ and $\hat{j}$.

Table 2.14 shows the speaker classification accuracy when S1 is correctly classified at SNR's of -10, -5, 0, 5 and 10 dB for P=5,15,25 and All (no secondary classifier). We have chosen SNR's symmetric about 0 dB as both the sources are speakers and in realistic cases, S1 becomes the dominant speaker when S1S2R is positive and vice versa. In the case of speaker classification on noisy speech (Sec.2.8.1.2), we have seen that using a subset of atoms/frames gives better classification accuracy. But in the case of overlapped speech, we are simultaneously classifying both the speakers instead of hierarchical classification as in noisy speech. So, we expect better classification using all the frames/atoms in overlapped speech since the assumption that low energy frames may contain only noise does not hold here. So, we compare the classification accuracy using all the frames with that of using the subset of 60% high energy features and 80% of high energy atoms in Table 2.14. It is seen that we get better accuracy without using any secondary classifier at low S1S2R and all atoms/frames. Using P=5 gives better accuracy of 100% using K-medoid dictionary. Using K-medoid dictionary with all

the frames/atoms gives better accuracy than TDCS-0.8 and subset of frames mostly. Using a subset of frames/atoms gives better accuracy at high S1S2R when the energy of speaker S1 dominates.

Table 2.14: Average S1 speaker classification accuracy as a function of P (the No. of top entries P used in the secondary classifier) using ASNA-L1. The results are shown for using all atoms/features (All) and a subset of 60% high energy features and 80% of high energy atoms (Subset).

S1S2R		-10 dB		-5 dB		0 dB		5 dB		10 dB	
		All	Subset	All	Subset	All	Subset	All	Subset	All	Subset
P=5	TDCS-0.8	37.93	24.14	51.72	51.72	68.97	65.52	86.21	86.21	96.55	100.00
	K-medoid	31.03	17.24	68.97	55.17	79.31	65.52	93.10	93.10	100.00	100.00
P=15	TDCS-0.8	55.17	20.69	68.97	37.93	82.76	65.52	89.66	89.66	89.66	96.55
	K-medoid	44.83	24.14	72.41	51.72	79.31	75.86	89.66	89.66	93.10	96.55
P=25	TDCS-0.8	44.83	17.24	65.52	41.38	79.31	68.97	93.10	93.10	93.10	96.55
	K-medoid	51.72	20.69	72.41	48.28	82.76	72.41	93.10	89.66	93.10	96.55
All	TDCS-0.8	48.28	17.24	65.52	41.38	79.31	65.52	93.10	93.10	93.10	96.55
	K-medoid	48.28	20.69	68.97	48.28	86.21	68.97	93.10	89.66	93.10	96.55

Table 2.15 shows the speaker classification accuracy when S2 is correctly classified at SNR of -10, -5, 0, 5 and 10 dB for P=5,15,25 and All (no secondary classifier). The results are complimentary to Table 2.14 as the accuracy decreases with increase in S1S2R.

Table 2.15: Average S2 speaker classification accuracy as a function of P. The results are shown for using all atoms/features (All) and a subset of 60% high energy features and 80% of high energy atoms (Subset).

S1S2R		-10 dB		-5 dB		0 dB		5 dB		10 dB	
		All	Subset	All	Subset	All	Subset	All	Subset	All	Subset
P=5	TDCS-0.8	96.55	89.66	82.76	89.66	68.97	68.97	51.72	37.93	17.24	13.79
	K-medoid	100.00	93.10	93.10	89.66	72.41	75.86	55.17	51.72	20.69	17.24
P=15	TDCS-0.8	96.55	96.55	89.66	93.10	86.21	75.86	79.31	37.93	51.72	10.34
	K-medoid	96.55	93.10	89.66	86.21	82.76	82.76	75.86	51.72	58.62	20.69
P=25	TDCS-0.8	89.66	93.10	86.21	89.66	82.76	75.86	62.07	31.03	37.93	10.34
	K-medoid	93.10	96.55	89.66	89.66	82.76	86.21	72.41	37.93	31.03	17.24
All	TDCS-0.8	89.66	93.10	86.21	89.66	82.76	75.86	55.17	27.59	31.03	10.34
	K-medoid	93.10	96.55	89.66	89.66	82.76	86.21	58.62	41.38	24.14	17.24

Table 2.16 shows the speaker classification accuracy when both S1 and S2 are correctly classified in the same overlapped speech at SNR of -10, -5, 0, 5 and 10 dB for P=5,15,25 and All (no secondary classifier). It is seen that we get the best accuracies at S1S2R of 0 dB, since both the speakers have equal energy and are more likely to get correctly classified in the same utterance.

Table 2.17 shows the percentage of times 1st choice based on highest TSW is correctly classified as one of the ground truth speakers in the overlapped speech. It is seen that using K-medoid dictionary without any secondary classifier and all frames/atoms give 100% consistently except at S1S2R of 10 dB.

Table 2.18 shows the percentage of times 2nd choice based on second highest TSW is correctly classified in the overlapped speech. It is seen that we get the highest value using K-medoid dictionary

Table 2.16: Average speaker classification accuracy if the both the speakers are correctly classified in the same overlapped speech as a function of P. The results are shown for using all atoms/features (All) and a subset of 60% high energy features and 80% of high energy atoms (Subset).

S1S2R		-10 dB		-5 dB		0 dB		5 dB		10 dB	
		All	Subset	All	Subset	All	Subset	All	Subset	All	Subset
P=5	TDCS-0.8	34.48	17.24	37.93	41.38	41.38	37.93	44.83	27.59	17.24	13.79
	K-medoid	31.03	10.34	62.07	44.83	55.17	44.83	51.72	44.83	20.69	17.24
P=15	TDCS-0.8	51.72	17.24	58.62	34.48	68.97	48.28	68.97	34.48	44.83	10.34
	K-medoid	41.38	20.69	62.07	37.93	62.07	58.62	65.52	41.38	51.72	20.69
P=25	TDCS-0.8	34.48	13.79	51.72	37.93	62.07	48.28	55.17	27.59	34.48	10.34
	K-medoid	44.83	17.24	62.07	37.93	65.52	58.62	65.52	34.48	24.14	17.24
All	TDCS-0.8	37.93	13.79	51.72	37.93	62.07	44.83	48.28	24.14	27.59	10.34
	K-medoid	41.38	17.24	58.62	37.93	68.97	55.17	51.72	37.93	20.69	17.24

Table 2.17: Percentage of times 1st choice based on the highest TSW is correctly classified in the overlapped speech as a function of P. The results are shown for using all atoms/features (All) and a subset of 60% high energy features and 80% of high energy atoms (Subset).

S1S2R		-10 dB		-5 dB		0 dB		5 dB		10 dB	
		All	Subset	All	Subset	All	Subset	All	Subset	All	Subset
P=5	TDCS-0.8	96.55	96.55	82.76	82.76	82.76	82.76	89.66	82.76	93.10	89.66
	K-medoid	96.55	96.55	89.66	82.76	89.66	86.21	89.66	96.55	96.55	89.66
P=15	TDCS-0.8	96.55	89.66	96.55	82.76	96.55	79.31	89.66	75.86	89.66	82.76
	K-medoid	100.00	93.10	100.00	96.55	100.00	96.55	96.55	100.00	100.00	93.10
P=25	TDCS-0.8	96.55	93.10	100.00	86.21	96.55	79.31	89.66	82.76	93.10	86.21
	K-medoid	100.00	96.55	100.00	100.00	100.00	96.55	100.00	93.10	96.55	89.66
All	TDCS-0.8	93.10	93.10	96.55	86.21	96.55	79.31	89.66	82.76	93.10	86.21
	K-medoid	100.00	96.55	100.00	100.00	100.00	96.55	100.00	93.10	96.55	89.66

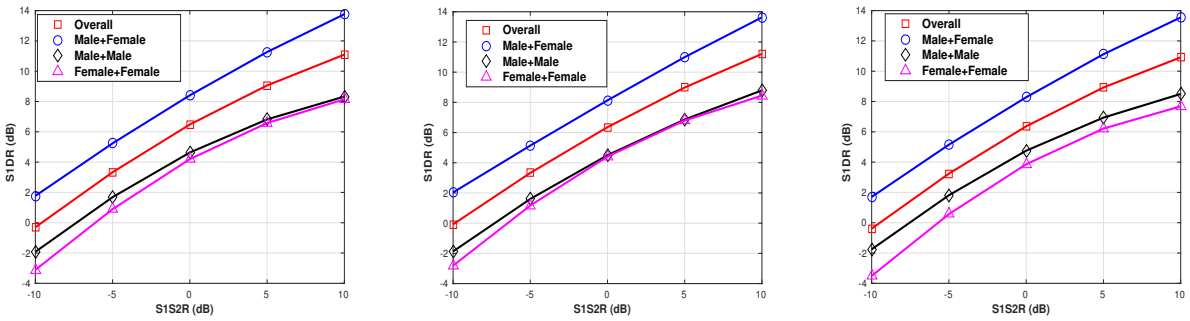
at S1S2R of 0 dB.

Table 2.18: Percentage of times 2nd choice based on the highest TSW is correctly classified in the same overlapped speech as a function of P.

S1S2R		-10 dB		-5 dB		0 dB		5 dB		10 dB	
		All	Subset	All	Subset	All	Subset	All	Subset	All	Subset
P=5	TDCS-0.8	37.93	17.24	51.72	58.62	55.17	51.72	48.28	41.38	20.69	24.14
	K-medoid	34.48	13.79	72.41	62.07	62.07	55.17	58.62	48.28	24.14	27.59
P=15	TDCS-0.8	55.17	27.59	62.07	48.28	72.41	62.07	79.31	51.72	51.72	24.14
	K-medoid	41.38	24.14	62.07	41.38	62.07	62.07	68.97	41.38	51.72	24.14
P=25	TDCS-0.8	37.93	17.24	51.72	44.83	65.52	65.52	65.52	41.38	37.93	20.69
	K-medoid	44.83	20.69	62.07	37.93	65.52	62.07	65.52	34.48	27.59	24.14
All	TDCS-0.8	44.83	17.24	55.17	44.83	65.52	62.07	58.62	37.93	31.03	20.69
	K-medoid	41.38	20.69	58.62	37.93	68.97	58.62	51.72	37.93	20.69	24.14

Figure 2.25 shows the variation of S1DR with S1S2R using the ground truth S1 and S2 dictionaries for all, male+female, male+male and female+female combinations (as listed in the beginning of this section) of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF

recovery. It is observed that we get the best S1DR of 8.41 dB at S1S2R of 0 dB for male+female combination using TDCS-0.8 and ASNA-L1 recovery, while we get an overall S1DR of 6.48 dB, and 4.63 dB and 4.61 dB for male+male and female+female combinations, respectively. It is because the separability between male and female speakers is higher than between same gender due to the high difference in their pitch frequencies and harmonics. The S1DR values are slightly lower for other combinations of dictionaries and source recovery. Figure 2.25(b) shows the performance of S1DR using SNMF dictionary and sup.NMF recovery. Figure 2.26 shows similar plots by using the S1 and S2 dictionaries corresponding to the estimates of $\hat{i}$ and $\hat{j}$. It is seen that S1DR on male+female combination decreases as S1S2R increases from 5 to 10 dB since the S2 classification accuracy decreases with increasing SNR and we get a wrong estimate of S2 index, $\hat{j}$. Figure 2.27 shows the variation of S2DR and it is seen that it decreases with increasing S1S2R since the relative energy of S2 with respect to S1 decreases.



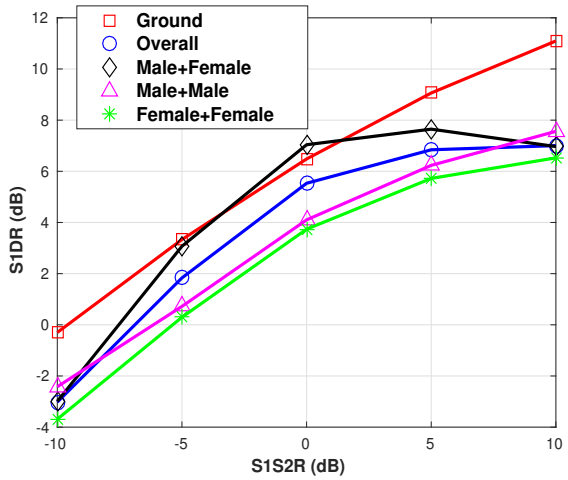
(a) TDCS-0.8 and ASNA-L1 (b) sup.NMF (c) K-medoid and ASNA-L1

Figure 2.25: Comparison of S1DR using ground truth S1 and S2 dictionaries for all, male+female, male+male and female+female combinations of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF recovery

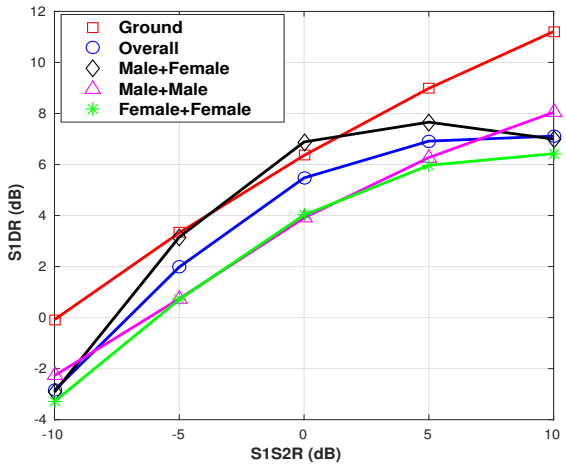
Figures 2.28 and 2.29 show the mean absolute error (MAE-S1S2R) and standard deviation of error in estimated S1S2R (STD-S1S2R) as a function of original S1S2R. It is seen that we get the least value of MAE-S1S2R at 0 dB S1S2R, for male+female combination due to high separability between male and female speakers.

2.10 Perceptual evaluation of source separation

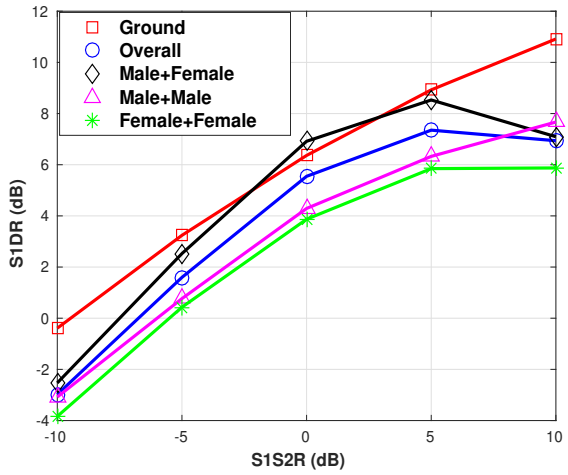
Most of the hearing aids do amplification of the sounds that reach the ear, while some of them also compensate (equalize) for the specific auditory frequency response of the person with hearing disability (PWHD). Hence, when the speech contains a significant amount of background noise, the perception is further affected. While this issue is true also for the people with normal hearing, the effect is more pronounced for PWHD. Thus, a system is highly desirable, which possesses the capability of separating the speech from the noise. We have reported our work on a system that identifies the noise type (among a set of 15 expected noise sources) as well as the speaker (from a known set of speakers



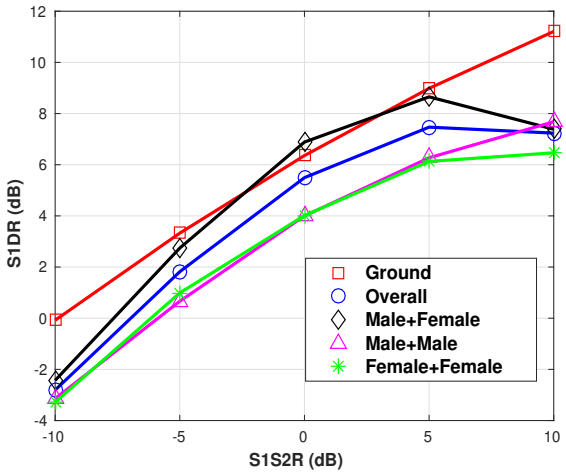
(a) TDCS-0.8 and ASNA-L1



(b) TDCS-0.8 and sup.NMF



(c) K-medoid and ASNA-L1

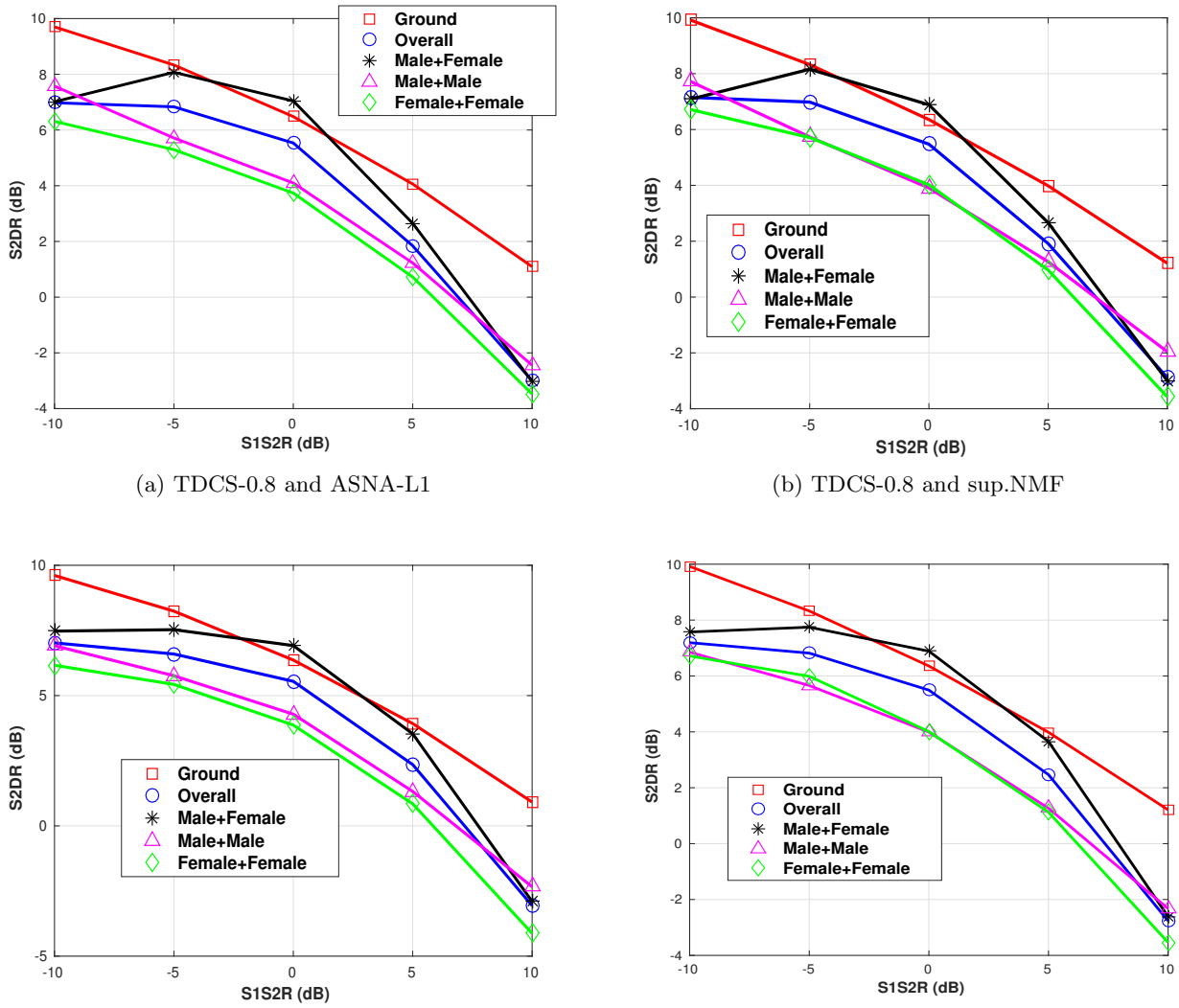


(d) K-medoid and sup.NMF

Figure 2.26: Comparison of S1DR using S1 and S2 dictionaries after classification for all, male+female, male+male and female+female combinations of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF recovery

that the person normally meets with during his daily life) and then extracts the speech from the noisy speech in Sec.2.8. We validate the effectiveness of the system by 69 human evaluators. The subjective evaluation indicates a significant enhancement in the perceived quality of speech and the information communicated.

For perceptual evaluation of the enhanced speech after separation, dictionaries of size 1000 atoms are learnt using random selection of mag.STFT features. The speaker sources are taken from TIMIT and CMU KED English, MILE Kannada, Tamil and English databases, IIIT Blizzard Telugu and Hindi databases. Noise sources are taken from factory, babble, jet cockpit from NOISEX database,



(a) TDCS-0.8 and ASNA-L1

(b) TDCS-0.8 and sup.NMF

(c) K-medoid and ASNA-L1

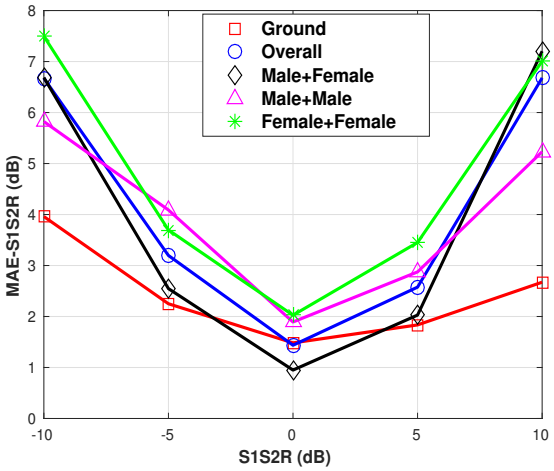
(d) K-medoid and sup.NMF

Figure 2.27: Comparison of S2DR using S1 and S2 dictionaries after classification for all, male+female, male+male and female+female combinations of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF recovery

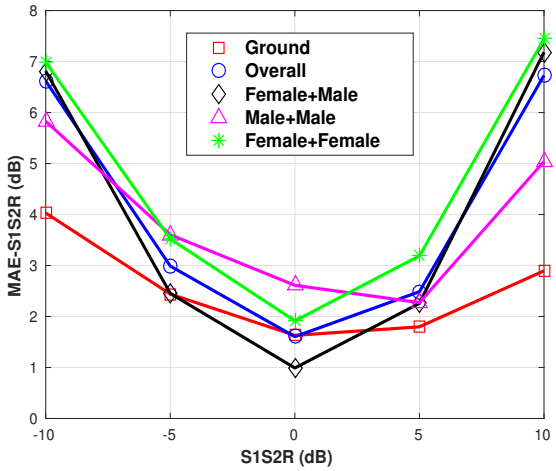
Traffic and Veena noise downloaded from online [96]. The speech and noise are mixed at an SNR of 0 dB to simulate the noisy speech.

The enhanced utterances are perceptually evaluated. The total No. of human evaluators are 69 with 56 male and 13 female listeners. The native language of the listeners is distributed among Kannada, Tamil, Telugu, Hindi, Malayalam, Marathi and Gujarathi. For 2 out of 69 listeners, perceptual evaluation has been done on speech mixed with speech from other speaker (overlapped speech).

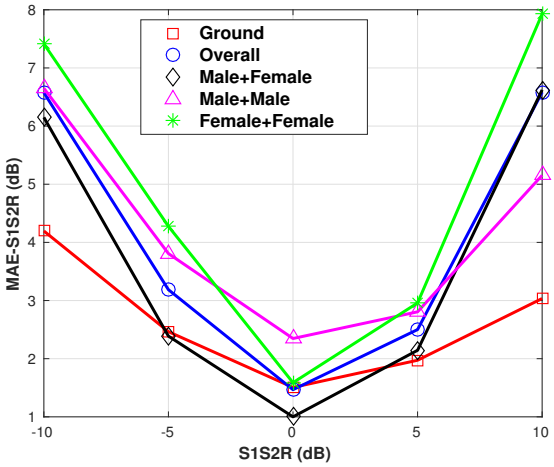
The listening test was done in an open room with some ambient noise using speakers and headphones. The evaluators are required to listen to the noisy speech and the enhanced speech after



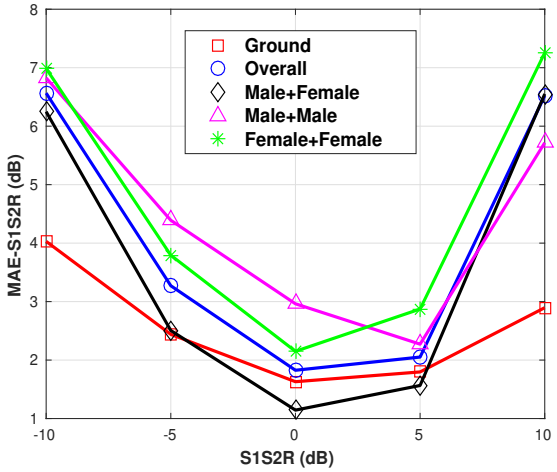
(a) TDCS-0.8 and ASNA-L1



(b) TDCS-0.8 and sup.NMF



(c) K-medoid and ASNA-L1

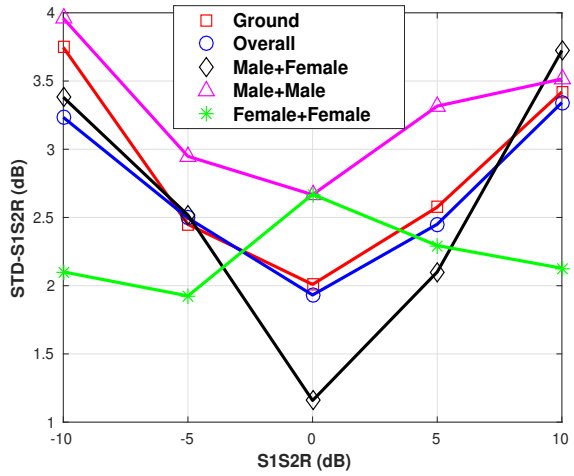


(d) K-medoid and sup.NMF

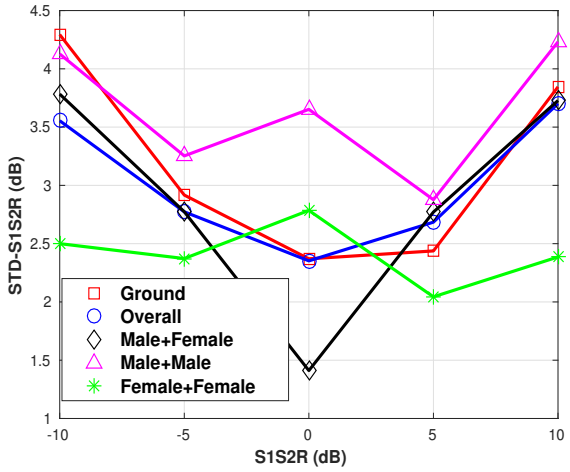
Figure 2.28: Comparison of MAE-SNR using S1 and S2 dictionaries after classification for all, male+female, male+male and female+female combinations of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF recovery

separation. A subjective score (opinion score) ranging from 1 to 5, where 1 indicates unsatisfactory speech quality and annoying and objectionable distortion while 5 indicates excellent speech quality and imperceptible distortion [97] is used for evaluation on the enhanced speech.

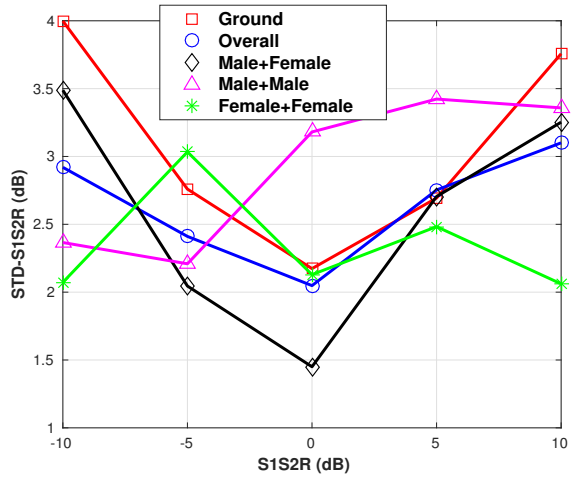
Figure 2.30 shows the distribution of the subjective scores and the age of the 69 evaluators. The mean opinion score (MOS) is 4.03 with a standard deviation of 0.68 (shown in Table 2.19) which shows the efficacy of the ASNA-L1 algorithm for separation. Around 58% of the evaluators have given a score of 4 while 23% have given a score of 5. The scores are better for overlapped speech of a male with a female speaker over that containing both male speakers.



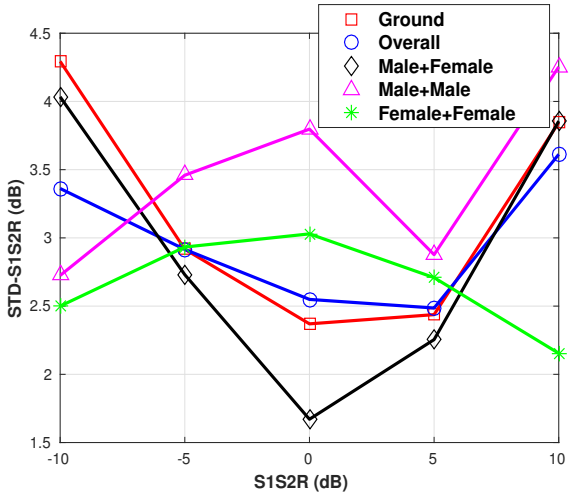
(a) TDCS-0.8 and ASNA-L1



(b) TDCS-0.8 and sup.NMF



(c) K-medoid and ASNA-L1



(d) K-medoid and sup.NMF

Figure 2.29: Comparison of STD-SNR using S1 and S2 dictionaries after classification for all, male+female, male+male and female+female combinations of overlapped speech using TDCS-0.8 and K-medoid dictionaries with ASNA-L1 and sup.NMF recovery

2.11 Noise classification and segmentation in a noisy conversation

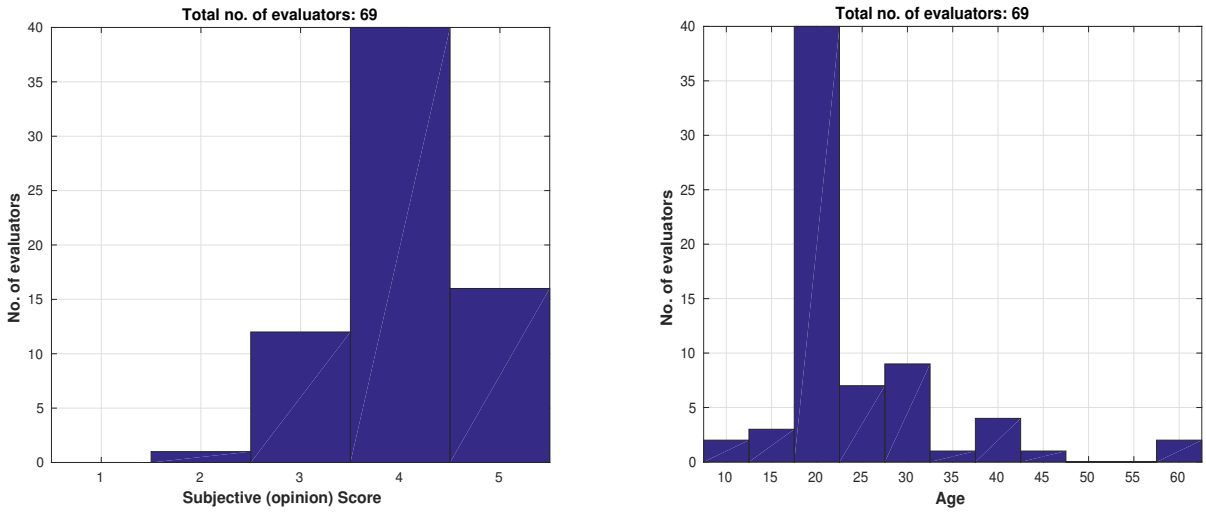
In this section, we deal with a noisy conversation between two people. A noisy conversation can be seen as a concatenation of two noisy speech signals, each of them as given in Sec.2.8.

The test signal, $y[n]$ is simulated as the concatenation of two noisy speech signals, $y^1[n]$ and $y^2[n]$. Each $y^l[n], l \in 1, 2$ is simulated as a linear combination of a speech, $y_{sp}^{il}[n]$ and a noise source, $y_{ns}^{jl}[n]$ as

$$y^l[n] = y_{sp}^{il}[n] + y_{ns}^{jl}[n] \tag{2.24}$$

Table 2.19: Mean opinion score and its standard deviation over 69 evaluators

MOS score	Standard deviation
4.03	0.68



(a) Distribution of subjective score

(b) Distribution of age

Figure 2.30: Histogram of (a) the subjective score given by 69 evaluators, and (b) The age distribution of the human evaluators.

Both the speaker and noise sources for the two noisy speech signals, $y^1[n]$ and $y^2[n]$ are different and are constrained to belong to a specific set of speaker and noise sources. So, $y[n]$ contains two different speakers and noise sources, each noise segment containing speech utterance by a single speaker. The first noise segment contains a female speaker and the second, a male speaker. This might be considered as a simulation of a telephonic conversation, where each speaker speaks in a different noise environment.

Figure 2.31 shows clean speech utterances from two different speakers concatenated without any gap in between them in (a), concatenated noise sources in (b), and the noisy conversation in (c) as a linear combination of (a) and (b) at an SNR of 0 dB. The original (4364.87 ms) and estimated (4365 ms) instants of transition from babble to pink noise are depicted in the figure.

Given the simulated noisy conversation, we find the instant of transition from one noise to another and classify the noises. Within each noise segment, we can estimate the speaker source and separate the speech from the noise, since it reduces to the problem of noisy speech containing a single speaker and noise source as in Sec.2.8.

2.11.1 Transition detection and classification of noises

Initially, the noise class is assigned frame-wise to a subset of features selected based on cosine-similarity measure similar to the procedure for noise classification in Sec.2.8.1.1. The steps for noise classification using this approach are enumerated in Algorithm 7. The estimated noise classes are $\hat{j}_1$, $\hat{j}_2$ based on

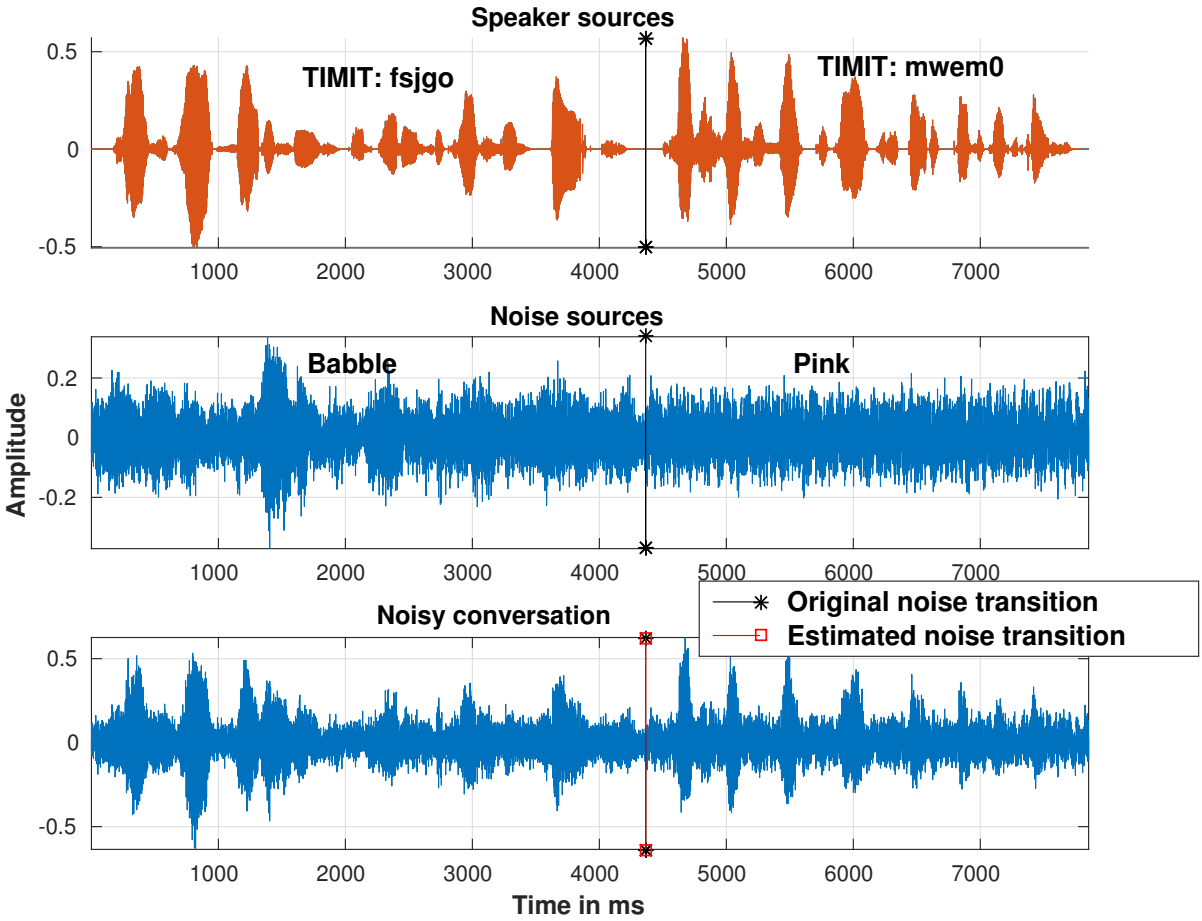


Figure 2.31: Illustration of two speaker sources, two noise sources and the noisy speech signal at an SNR of 0 dB.

the top two $TSDR$ as in Algorithm 3. The sequence of occurrence of the estimated noise indices $\hat{j}_1$, $\hat{j}_2$ is determined based on the centroids of the frame indices corresponding to $\hat{j}_1$, $\hat{j}_2$. The approximate transition frame is shown as that frame for which the difference of the No. of indices upto that frame (within the subset of frames) corresponding to $\hat{j}_1$ and $\hat{j}_2$ is maximum, as intuitively this difference is maximum at the transition frame. The exact transition instant is arrived at as the mid-point of the approximate transition frame.

Algorithm 7 *Noise segmentation and classification*

1: n is the total No. of features(frames) in the test conversation
2: Select a subset of features as mentioned in Algorithm 3, from the n mag.STFT features extracted from the noisy conversation. $\mathbf{ind}_{n_s}$ contains the indices of the subset

3: Do frame-wise assignment of noise classes as $\hat{\mathbf{J}}$ on the subset of features, $\mathbf{ind}_{n_s}$ using SDR measure and ASNA recovery as explained in Sec.2.6.1

4: Assuming only two noise classes are present within the whole test signal, pick two noise classes $\hat{j}_1, \hat{j}_2$ based on the top two $TSDR$ as in Algorithm 3.

5: **Noise classification:** The estimated noise classes are $\hat{j}_1, \hat{j}_2$. The first and second noise classes are decided based on the centroids of the subsets of the indices $\mathbf{ind}_{n_s}$ assigned to $\hat{j}_1$ and $\hat{j}_2$ respectively.

6: **Noise transition:** For all $1 \leq p_t \leq n$, find the difference between the No. of frames in the subset $\mathbf{ind}_{n_s} \cap \{1, 2, ...p_t\}$ assigned to estimated first noise class $\hat{j}_1$ and the No. of frames in $\mathbf{ind}_{n_s} \cap \{1, 2, ...p_t\}$ assigned to estimated second noise class $\hat{j}_2$.

7: Find the p_t which gives the maximum difference in the above step and assign it as the estimated noise transition frame, tr_{fr}.

8: The estimated noise transition instant is obtained as the mid-point of the frame tr_{fr} as tr_{est}.

end

Although the above algorithm assumes that the signal consists of two consecutive noise segments, it can be generalized to the segmentation of a signal containing multiple noises and transitions. In a test noisy conversation, there may be multiple transitions and an unknown No. of segments corresponding to different noise sources. We assume the noise source is the same across various frames over a significant span of time since the noise class may not change unless the speaker is traveling. So, we follow a top-down approach of classifying and segmenting a test utterance. For practical purposes, block-wise classification is more realistic than frame-wise classification, as seen earlier. Initially, the noise class is assigned frame-wise to a subset of features as in Algorithm 7. Then, assuming a constant noise source across the whole test signal, and then recursively dividing the test utterance into two segments, segment level classification is performed. A valid assumption is made that each segment has a uniform noise present for at least 2 seconds. Algorithm 8 is the generalization of Algorithm 7 for multiple transitions and noise sources. Here, we estimate the No. of noise sources, No. of segments having same noise across the segment, transition instants across various segments and the noise label corresponding to each segment. A divide and conquer approach is proposed by recursively dividing the test frames into two equal parts until 90% of the subset of frames within each part are classified as the same class or the number of component frames corresponds to less than 2 seconds. In the case where 90% of the subset of frames within a part are not classified as the same class, the maximum occurring noise class across the subset of features within the segment defines the noise label of the segment. Segments which do not contain any frames within ind_{ns} are still labelled as null. The class label of these frames is assigned as the class label corresponding to the nearest centroid of the indices of N_{seg} segments. The total No. of noise segments are N_{seg} and the corresponding labels are assigned sequentially as $\hat{j}_1 ... \hat{j}_{N_{seg}}$ based on the locations of the centroids. The noise transition instants are

assigned by considering two consecutive noise segments as the noisy speech signal used in Algorithm 7 and following steps 6-8.

So, Algorithm 8 estimates the unknown No. of noise segments as N_{seg}, the noise labels $\hat{j}_1...\hat{j}_{N_{seg}}$, and the noise transition instants $tr_{est_1}...tr_{est_{N_{seg}-1}}$. It is to be noted that total No. of unique noise classes may be less than N_{seg} as two noise segments belonging to same class may occur as non-consecutive segments.

The computational complexity of Algorithm 7, 8 is mainly due to the framewise noise classification stage (same as Algorithm 3) and the noise segmentation stage. Algorithm 7 takes $O(n)$ and Algorithm 8 takes $O(n \log n)$ (due to DivideRecursive function) for n frames for the noise segmentation stage.

Algorithm 8 *Generalized noise segmentation and classification*

1: n is the total No. of features(frames) in the test conversation
2: Initialize the class labels of all the test frames as null, $\mathbf{L} = null$
3: Select a subset of features as mentioned in Algorithm 3, from the n mag.STFT features extracted from the noisy conversation. $\mathbf{ind}_{n_s}$ contains the indices of the subset
4: Do frame-wise assignment of noise classes as $\hat{\mathbf{J}}$ on the subset of features, $\mathbf{ind}_{n_s}$ using SDR measure and ASNA recovery as explained in Sec.2.6.1
5: DIVIDERECURSIVE($\hat{\mathbf{J}}, 1, n, \mathbf{ind}_{n_s}$)
6: **No. of noise segments** The No. of segments are N_{seg} (distinct segments identified in the call to DivideRecursive function), each segment labelled as $\hat{j}_1..\hat{j}_{N_{seg}}$ in the previous step
7: Find the centroid of the indices $\mathbf{ind}_{n_s}$ assigned to $\hat{j}_1...\hat{j}_{N_{seg}}$ within the segments of the corresponding labels as $cent_{\hat{j}_1}...\ cent_{\hat{j}_{N_{seg}}}$
8: **for** All the n frames ($p = 1...n$) **do**
9: **if** $\mathbf{L}(p) = null$ **then**
10: Find the absolute differences between the centroid of the present frame, $cent_{frm}$ and $cent_{\hat{j}_1}...cent_{\hat{j}_{N_{seg}}}$ as $dist_{\hat{j}_1}...dist_{\hat{j}_{N_{seg}}}$
11: Update the class label $\mathbf{L}(p)$ of the present frame as the index corresponding to the minimum of $dist_{\hat{j}_1}...dist_{\hat{j}_{N_{seg}}}$
12: **end if**
13: **end for**
14: **Noise classes:** The estimated noise classes for N_{seg} segments are $\hat{j}_1...\hat{j}_{N_{seg}}$
15: **Noise transitions:** For all successive pairs of segments with labels $\hat{j}_i, \hat{j}_{i+1}$, identify the noise transition instants as in steps 6,7,8 in Algorithm 7, considering each pair of segments as a noisy conversation dealt with in Algorithm 7
16: For $N_{seg} - 1$ pairs of noise segments, we get $tr_{est_1}...tr_{est_{N_{seg}-1}}$ transition instants
17: **function** DIVIDERECURSIVE($\hat{\mathbf{J}}, a, b, \mathbf{ind}_{n_s}$)
18: Select a subset of the indices as $\hat{\mathbf{J}}_{sel} = \hat{\mathbf{J}}\ \forall\ \mathbf{ind}_{n_s} \in a, a+1, ..., b-1, b$
19: **if** $\hat{\mathbf{J}}_{sel}$ is empty **then return**

end if
Find the percentage occurrence p_k of each noise index in $\hat{\mathbf{J}}_{sel}$
If the percentage occurrence of any noise class k is greater than 90% i.e.
if $p_k \geq 90$ **then**
Update the class labels of the entire segment as $\mathbf{L}(a:b) = k$ **return**
else if $(b-a) \leq nf_2$ (nf_2 is number of frames corresponding to 2 seconds) **then**
Find the noise index k having the maximum occurrence across $\hat{\mathbf{J}}_{sel}$
Update the estimate of the class labels as $\mathbf{L}(a:b) = k$ **return**
end if
Assign $m = \dfrac{a+b}{2}$
DIVIDERECURSIVE($\hat{\mathbf{J}}, a, m, \mathbf{ind}_{n_s}$)
DIVIDERECURSIVE($\hat{\mathbf{J}}, m+1, b, \mathbf{ind}_{n_s}$)
end function

end

2.11.2 Results on transition detection and noise classification

A noisy conversation is simulated by concatenating an utterance from a female speaker mixed with first noise source to that from a male speaker mixed with second noise source at SNR's of -10, 0, 10 and 20 dB. However, our approach is independent of the genders of the speakers used. The time gap between speech utterances is varied as no gap, 0.1 sec and 1 sec such that the transition occurs near the mid-point of the gap.

The results in this section are obtained using Algorithm 7 as concatenation of only two noise sources is considered. The Algorithm 8 is proposed as a future work to be explored and the results have not been shown.

Two different noise sources are randomly chosen for the first and the second parts such that all the noise sources are used in each part. The speaker classes are selected such that all the fifteen combinations are different and all female/male speakers are used. So, 15×15 combinations of noisy conversations are used for testing at different SNR's.

Table 2.20 shows the overall noise classification accuracy for both the parts in the noisy conversation at zero (no gap), 0.1 sec and 1 sec time gap using K-medoid and TDCS-0.9 dictionaries. We get 100% noise classification accuracy at -10 and 0 dB using K-medoid at 0.1 and 1 sec gaps. Noisy conversations having higher time gap have more noise-only frames and give better classification and detection of transitions.

Tables 2.21 and 2.22 show the mean absolute (MAE) and standard deviation (STD) of error (in milliseconds) in the detection of noise transition instant for various SNR's. We get the highest mean absolute error of 169.3 ms using TDCS-0.9 and no gap at 20 dB SNR. The high MAE is in the case of machinegun noise which gets mostly misclassified at high SNR. If we exclude machinegun noise, we

Table 2.20: Noise classification accuracy as a function of input SNR at various time gaps between utterances.

SNR	Time gap	-10 dB	0 dB	10 dB	20 dB
TDCS-0.9	No gap	100.00	98.89	98.67	95.78
	0.1 sec	100.00	99.33	98.22	95.33
	1 sec	100.00	100.00	99.11	97.78
K-medoid	No gap	100.00	99.56	97.33	95.33
	0.1 sec	100.00	100.00	97.11	94.89
	1 sec	100.00	100.00	99.33	97.78

get a MAE of 42.74 ms. As the time gap increases, the MAE and STD decrease due to the increase in the No. of noise-only frames, and we are able to detect the transition instant more accurately.

Table 2.21: Mean absolute error (in msec) in the detection of noise transition instant, for various SNR's and intervals between concatenated utterances.

MAE	Time gap	-10 dB	0 dB	10 dB	20 dB
TDCS-0.9	No gap	16.05	68.35	98.58	169.30
	0.1 sec	10.37	46.54	96.50	112.27
	1 sec	9.84	9.84	38.64	21.90
K-medoid	No gap	14.07	47.41	102.68	146.07
	0.1 sec	11.08	32.19	114.77	145.72
	1 sec	10.14	10.14	38.94	22.54

2.12 Conclusion and future work

A new approach to noise and speaker classification has been proposed adopting ASNA as the source recovery algorithm. Experiments have shown a good overall frame level accuracy of 97.8% on noise classification. We plan to explore and devise other discriminative dictionary learning and source recovery algorithms for faster and more efficient source classification.

We have classified the speaker and noise sources from noisy speech signals with good accuracy using various simple dictionary learning methods and sparse representation. We have also estimated the SNR and reported the separation performance. A novel approach is proposed for the classification

Table 2.22: Standard deviation of error (in msec) in the detection of noise transition instant, for various SNR's and time gaps between utterances.

STD	Time gap	-10 dB	0 dB	10 dB	20 dB
TDCS-0.9	No gap	75.61	316.96	452.33	635.62
	0.1 sec	14.84	248.07	453.84	524.55
	1 sec	12.62	12.62	331.55	165.48
K-medoid	No gap	45.98	220.71	441.30	583.70
	0.1 sec	17.16	186.99	503.63	597.37
	1 sec	15.12	15.12	331.68	169.64

and separation of noisy speech signals commonly occurring in telephonic conversations. Since mobile communication is ubiquitous nowadays, our approach can be used for tracking the speaker/noise sources and noise adaptive speech enhancement using sparse representation based methods. We have shown how updation of dictionaries using parts of the test signal itself improves the classification and separation performance. As a future work, we plan to use machine learning techniques to learn discriminative dictionaries so as to classify multiple classes of noise and speech signals, and mixed audio signals. Use of discriminative dictionaries may classify the various components in a mixed signal like language, speaker, gender, music and noises in a more generic way.

Chapter 3

Relative occurrences of extrema for detection of transitions between broad phonetic classes

Detection of transitions between broad phonetic classes in a speech signal has applications such as landmark detection and segmentation. The proposed hierarchical method detects silence to non-silence transitions, sonorant to non-sonorant transitions and vice-versa. The subset of the extrema (minimum or maximum amplitude samples) above a threshold, occurring between every pair of successive zero-crossings, is selected from each frame of the bandpass filtered speech signal. Locations of the first and the last extrema lie on either side far away from the mid-point (reference) of a frame, if the speech signal belongs to a non-transition segment; else, one of these locations lie within a few samples from the reference, indicating a transition frame. The transitions are detected from the entire TIMIT database for clean speech and 93.6% of them are within a tolerance of 20 ms from the phone boundaries. Sonorant, unvoiced non-sonorant and silence classes and their respective onsets are detected with an accuracy of about 83.5% for the same tolerance with respect to the labelled TIMIT database as reference. The results are as good as, and in some respects better than the state-of-the-art methods for similar tasks. The proposed method is also tested on the test set of the TIMIT database for robustness with respect to white, babble and Schroeder noise, and about 90% of the transitions are detected within a tolerance of 20 ms at the signal to noise ratio of 5 dB .

3.1 Introduction

In the previous chapter, we addressed classification of noise and speaker sources, which can be seen as a higher level classification; and subsequent separation of speech and noise components. Sparse representation of the mag.STFT features using dictionaries learnt from the training database was used for the same. Non-negative dictionaries were learnt by constraining the number of atoms and the mutual/ cross coherence and other dictionary learning methods. Feature selection for classification

has been explored by dynamically selecting noise and speaker features from the test signal.

In this chapter, we look into the lower level classification of speech segments in an audio signal. We segment the speech signal into different regions by an acoustic phonetics knowledge based approach. First, we extract a novel set of temporal features from the speech and the bandpass filtered signal. We then use these features to derive a rule based algorithm for detecting the transitions and segmenting the speech signal into different classes.

3.1.1 Segmentation problem

During speech production, the articulators continually move resulting in a speech signal with almost continuous formant tracks. Also, the source process is influenced by the preceding or the succeeding phone, as for example the glottal abduction during a vowel-consonant transition, the presence of frication noise following a burst, or the presence of noise components at the onset of a vowel following a strong fricative. Thus, the adjacent phones have a considerable influence on the temporal and spectral properties of a short segment of speech corresponding to the so called current phone [98]. However, we perceive clean speech as if it is made up of a sequence of distinct sounds, thus evoking an expectation that the signal can be segmented into non-overlapping intervals corresponding to phones. Hence, speech segmentation is a challenging problem. Despite the lack of phone-wise segmentation property in a speech signal, there are clearly marked events or transitions or landmarks arising due to an abrupt change of source process (voiced/unvoiced) and/or an abrupt movement of an articulator (sudden release as in stops, switch over from oral to nasal output as for nasals). Detection of such events serves to guide semi-automatic segmentation, variable frame-rate analysis or analysis around landmarks to extract distinctive features (DFs) or manner classes [99] or phonetic features (PFs). Mesgarani et al. [100] used high density direct cortical surface recordings in humans and found response selectivity to distinct phonetic features in the superior temporal gyrus (STG), which shows acoustic-phonetic representation of speech in human STG.

3.1.2 Literature review

There are three broad approaches to segmentation: (i) sequential, non-overlapping segmentation based on phones, (ii) parallel, multiple segmentations based on DFs, and (iii) hierarchical segmentation based on PFs. Classification of a speech signal, phone-wise or feature-wise, can also be interpreted as performing segmentation, since it automatically divides the speech signal into distinct segments.

The first view, namely, the phone based segmentation, is motivated by the perception of speech as a series of distinct units. As per this view, any speech signal is a sequence of non-overlapping intervals, each representing a phone. This assumption is widely used in manual labeling of the speech databases. Such a labeling scheme contradicts the acoustic-phonetics knowledge that a given frame of speech signal is strongly influenced by the neighboring phones. However, it is understood that the manual labeling of phones must be considered along with the surrounding context of phones. In phone level segmentation, abrupt changes in the short-time spectra are marked as transition events [101, 102, 103, 104, 105, 106]. Various short-time spectral representations have been used: linear

prediction smoothed spectral envelope, ensemble interval histogram, auditory sub-band filter outputs, mel frequency cepstral coefficients (MFCCs), weighted MFCCs, etc. In addition to the standard Euclidean and Mahalanobis distance measures [101], cross-correlation of short-time spectra [102], model fitting [103], maximum likelihood estimates and template matching [104] have also been used to detect segment boundaries.

The need for segmentation is averted using statistical models, such as hidden Markov models, for the classification of phones [107]. It may appear that the statistical approach at once solves both the segmentation and classification problems. The disadvantages with the statistical method are that it requires a huge amount of labeled database for training and it is sensitive to the recording conditions and background noise. Any change in the recording condition may call for re-training. Although hidden Markov model based forced alignment [108] does not require a training database, it assumes that the phone sequence of the utterance to be segmented is known. However, in this work, we deal with segmentation when the phone sequence is unknown.

The superior speech perception performance of humans in degenerate conditions [109] points to the possibility that humans may utilize other sources of knowledge such as distinctive or phonetic features [100]. The importance of integration of phonetic knowledge in speech technology [110] has been discussed. Several studies have shown that the additional use of DFs or PFs improves the speech recognition performance of a hidden Markov model classifier [111, 112]. As PFs and DFs are useful for segmentation, it may be inferred that segmentation improves human perception of speech.

The second approach to segmentation is based on distinctive features, a view based on phonology. It is postulated that each speech sound is a bundle of (about 16) binary DFs [113],[114]. In this model, speech signal consists of a parallel stream of DFs. The presence of each of the DFs extends over different, overlapping intervals with their own boundaries. Acoustic description of DFs given by Jakobson et al. [113] has remained qualitative in nature since there is no robust automatic method to extract these descriptors. Chomksy et al. [114] have proposed articulation based DFs. In order to extract these DFs, King et al. [115] used frame-wise analysis with MFCCs and their derivatives (39 features) as the acoustic feature vector input to a neural network classifier trained for each DF separately. Though the frame-wise accuracy for the individual DF is high (>90%), the accuracy for the joint or simultaneous occurrence (all correct) of the DFs for a given phone is low (around 50%), nearly the same as the front-end accuracy of the early template approach. Hidden Markov models [116] and support vector machine [117] have also been used for the extraction of DFs. Bromberg et al. [118] experimented with a bank of acoustic features along with a bank of classification strategies to extract attributes. No specific feature or classifier uniformly gave a high accuracy. A multilayer perceptron classifier gave an equal error rate (EER) of 10% for 'strident vs silence' but a poor EER of 25% for the feature 'high'. A support vector machine classifier gave an EER of 6% for 'silence' but a poor 42% for the feature 'mid'.

The third approach, based on the phonetic features, has two models. One of the models is based on the manner and place classification of speech sounds, a view inspired by the process of speech

production. This is similar to the approach of DF but with multi-valued features and only two parallel streams (manner and place). In their work on DFs, King et al. [115] also reported on the identification of manner and place features. The reported frame-wise accuracy is about 90% for the individual features but the accuracy for all the features being simultaneously correct is only 50%. A later extension of this study attempted to incorporate mutual dependencies amongst DFs [119] and found a marginal improvement in the accuracy. Juneja et al. [120],[112] report manner class (silence, vowel, sonorant consonant, fricative and stop) segmentation accuracy of about 79% [112] on a part of the test set of the TIMIT database, using MFCCs as well as acoustic parameters and support vector machine classifier.

Assuming a bundle of 16 binary DFs, there are $2^{16} = 65536$ possible representations and with multi-valued DFs (six manner and eight place), there are $2^{14} = 16384$ possible representations, whereas the number of phones is only of the order of 50. This discrepancy arises because a large number of DFs are mutually exclusive. All the DFs may not be relevant for a given phone. For example, the feature 'high/low' is relevant only for the manner class 'vowels' and irrelevant for consonants and the place feature is irrelevant for vowels. This leads to the second hierarchical classifier model within the approach of phonetic features. Here, a tree structure is used, where each node represents a broad class, which is sub-divided into two finer classes [121],[122]. For example, the signal is initially divided into the two broad classes of speech and silence. Then the node 'speech' divides into two branches, namely sonorants and non-sonorants. Sonorants are divided into syllabic (vowels) and consonantal (voiced consonants other than stops). The non-sonorants are divided into continuants (fricatives) and interrupted (stops) and so on.

In a hierarchical scheme of PFs, the segmentation problem can be reduced to a set of binary decision making problems. Alternately, if the phone corresponding to the final terminal is determined, then all the higher level nodes leading to that branch are also determined. For example, if a stop is detected from a segment of a continuous speech signal, then that segment is automatically assigned all the higher level nodes, namely, interrupted, non-sonorant, speech. Hence, research at determining either the PFs or an individual phone, irrespective of context, is being pursued by various researchers: (i) onset of a vowel given the segment contains a consonant-vowel transition [123],[124] (ii) offset of a vowel given the segment contains a vowel-consonant transition [125] (iii) semi-vowels [126], laterals [127], nasals [122] (iv) fricatives [128] (v) stop bursts [129],[130] and (vi) trills [131].

An issue related to the extraction of phonetic features is the landmark detection, landmark being an important transition dividing a speech signal into certain broad segments [121],[132],[133]. Conventionally, speech analysis for the extraction of acoustic features is carried out frame-wise. But in this alternative approach, speech signal around the landmarks is analyzed to extract the acoustic features, which are subsequently given as an input to a classifier to determine either the phones or phonetic features. Liu [133] has used the change of energy, over six sub-band signals, between two frames spaced 50 ms apart, for detecting four broadly defined landmarks. Salomon et al. [132] have used a set of twelve temporal parameters to detect three landmarks as well as for manner classification. Reddy [134]

proposed the use of intensity differences (peaks and valleys) to detect certain broad classes of sounds for a limited vocabulary, speaker dependent task. Stevens [121] has observed that certain landmarks may be located based only on abrupt amplitude changes in a speech signal.

3.1.3 About this work

Speech signal is non-stationary and has rapid variation in the time domain. In this chapter, we segment the speech signal into various classes and assign transitions based on various measures extracted from the time domain signal followed by a rule-based approach. Four different measures are proposed for detecting transitions between broad phonetic classes in a speech signal based on abrupt amplitude changes in the bandpass filtered and the quantized speech signal.

A measure is defined on the quantized speech signal to detect transitions between very low amplitude or silence (S) and non-silence (N) segments. These S-segments could be stop closures, pauses or silence regions at the beginning and/or ending of an utterance.

We propose two other measures to detect the transitions between sonorant and non-sonorant segments and vice-versa. We make use of the fact that most sonorants have higher energy in the low frequencies, than other phone classes such as unvoiced fricatives, affricates and unvoiced stops. For this reason, we use a bandpass speech signal (60-340 Hz) for extracting temporal features. For a transition within a sonorant (vowel to voiced consonant or vice-versa), the amplitude of the bandpass filtered speech signal does not change appreciably. However, for a transition from a sonorant to any of the unvoiced consonants, the amplitude changes suddenly from a high to a low value across the transition. The converse is also true. Thus, by tracking the relative locations of extrema in successive closely spaced analysis frames, we can detect the transitions between high (H) and low (L) amplitude segments and hence the broad phonetic classes in a speech signal.

When the amplitude of the bandpass filtered signal in a frame is very low, any change in the relative amplitude level is not reliable. Thus, when the mean difference between extrema amplitudes in a frame is very low, any transition is ignored.

The above rationale for the selection of features and the proposed algorithm for detecting transitions is based on an expectation born of the acoustic-phonetic knowledge of the different classes of speech sounds.

Combining the above types of transitions, the speech signal is divided into the five broad homogeneous classes: silence (S), high (H), low (L), high-low (HL) and low-high (LH). Based on the homogeneous classes, the speech signal is classified into the broad phonetic classes of sonorants, non-sonorants and silence. The proposed method is validated using the TIMIT database under clean and noisy conditions. The accuracy of detection and the temporal accuracy of the onset of these classes are computed. The results are noted to be comparable to those of state-of-the-art methods.

3.2 Proposed temporal features

From the normalized speech signal, we derive temporal features, which are independent of the amplitude level (gain) of the signal. The parameters and thresholds used are derived from a development set based on intuition and statistical observation.

3.2.1 Pre-processing

The speech utterance $s[n]$ is normalized after removing the mean value as:

$$s_z[n] = s[n] - \frac{\sum_1^F s[n]}{F} \; ; s_N[n] = \frac{s_z[n]}{\max|s_z[n]|} \tag{3.1}$$

where F is the total number of samples in the utterance and $s_N[n]$ is the normalized speech utterance. Frames extracted from $s_N[n]$, using a uniform frame shift of 5 ms, are used for deriving the temporal features.

3.2.2 Silence Index

To detect silence zones, we define a new measure called silence index. The normalized speech signal of 16 bit resolution is quantized to 9 bits by shifting right by 7 bits and a staircase signal is obtained. The size of the analysis frame is 10 ms. Whenever there are a minimum of three successive samples having the same value and whose absolute values are up to a threshold (two times the quantization level, 2^7), these samples are counted for the calculation of SI.

Silence Index (SI) is a dimensionless ratio (between 0 and 1), defined as

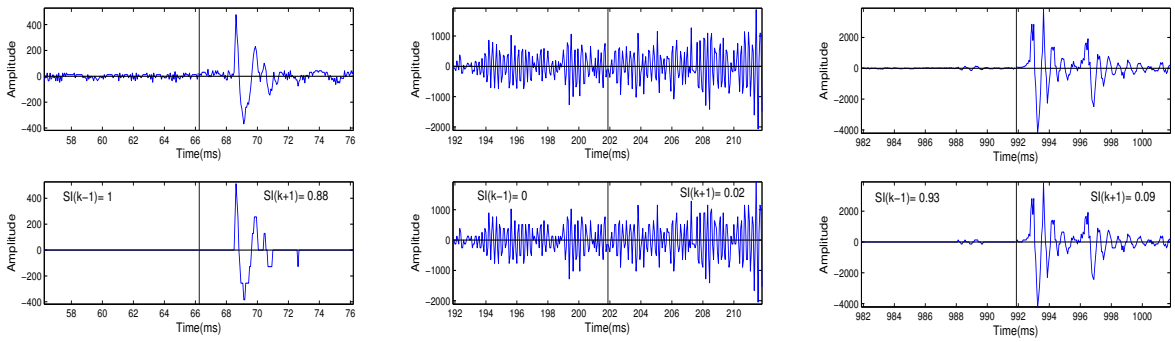
$$SI = \frac{count\ of\ samples\ below\ threshold\ having\ same\ value\ across\ 3\ succesive\ samples}{number\ of\ samples\ in\ the\ frame} \tag{3.2}$$

Since a frame shift of 5 ms is used, there is a new value of SI for every 5 ms segment.

Figure 3.1 shows the signal and its quantized counterpart, together with the SI values for three types of speech segments, each containing three overlapping frames: (a) silence containing a noisy impulse, (b) unvoiced and (c) a closure to burst transition segment. It may be noted that SI has a very high value, as desired for the silence segment, even in the presence of a large amplitude impulse. The SI is low for the unvoiced segment. During a closure-burst transition, there is a sharp decrease in the value of SI for two successive frames. We make use of such abrupt changes in the value of SI for detecting the transitions from/to silence segments.

3.2.3 Features based on the extrema in a frame

Features based on extrema are used to detect transitions from/to a sonorant segment. As sonorants have significant energy in the low frequencies below 500 Hz as compared to other segments, the normalized speech signal is bandpass filtered (BPF) using the following bell cosine shaped filter in the frequency domain:



(a) (b) (c)

Figure 3.1: Variation of silence index (SI) values with the variation in the nature of the signal across consecutive frames. Speech signal and the corresponding quantized signals for (a) Presence of a high amplitude pulse in a silence segment. (b) An unvoiced segment. (c) A stop closure to burst transition. (Note that the y-scales are different for the three plots)

$$h_B[f] = \left\{ \begin{array}{ll} 0.5 - 0.5\cos\left(\pi\left(\frac{f-f_1/2}{f_1/2}\right)\right), & f_1/2 \leq f < f_1 \\ 1 \qquad\qquad\qquad\qquad , & f_1 \leq f \leq f_2 \\ 0.5 + 0.5\cos\left(\pi\left(\frac{f-f_2}{f_2}\right)\right), & f_2 < f \leq 2f_2 \\ 0, \quad elsewhere & \end{array} \right\}$$

$f_1 = 70$ Hz, $f_2 = 250$ Hz and $h_B[f]$ is the frequency response of the filter. This filter has a cosine rising function from 35 to 70 Hz, unit gain from 70 to 250 Hz and cosine falling function from 250 to 500 Hz. The 3 dB frequencies of the bandpass filter are 60 and 340 Hz. This is close to 'Band 1' used by Liu [133] for landmark detection. The corresponding bandpass filtered signal $s_B[n]$ is given by:

$$s_B[n] = s_N[n] * h_B[n] \tag{3.3}$$

where $h_B[n]$ is the impulse response corresponding to $h_B[f]$. The BPF signal $s_B[n]$ is analyzed with a frame size of 40 ms, twice the pitch period corresponding to the assured minimum value of fundamental frequency of 50 Hz.

3.2.3.1 Selection of extrema based on a dynamic two-pass threshold

Let s_B^j denote the bandpass filtered signal between the first and the last zero crossings in the j^{th} frame. We define features based only on those extrema in s_B^j remaining after a 2-pass, frame adaptive threshold.

The first-pass positive threshold T_{P1}^j is defined as

$$T_{P1}^j = 0.5 \times mean(\{s_B^j[n]\})\ \forall s_B^j[n] > 0, \tag{3.4}$$

From s_B^j, all the positive peaks p_B^j between successive zero crossings are obtained. A subset of these peaks is selected as:

$$p_{B1}^j = \{p_B^j,\ \forall\ p_B^j > T_{P1}^j\} \tag{3.5}$$

The second-pass positive threshold, T_{P2}^j is defined as

$$T_{P2}^j = 0.5 \times mean\{p_{B1}^j\} \tag{3.6}$$

The set of peaks after the second pass is obtained as

$$p_{B2}^j = \{p_{B1}^j,\ \forall\ p_{B1}^j \geq T_{P2}^j\} \tag{3.7}$$

The factor of 0.5 is applied to compute the threshold based on the intuition that in the case of a midway transition, half of the peaks p_{B1}^j might lie below and the rest above T_{B2}^j . Similarly, from the valleys (negative peaks) between successive zero crossings, the set of valleys v_{B2}^j is obtained. Figure 3.2(a) shows a segment of voiced frame /ae/, its corresponding BPF output and the first zero crossing, the first and second-pass positive thresholds. It is to be noted that the peaks obtained after the first and second-pass are the same for this non-transition frame.

3.2.3.2 Relative occurrences of first and last extrema in a frame

The time of occurrence of the first extremum (OFE) and the last extremum (OLE) in p_{B2}^j or v_{B2}^j are measured with respect to the mid-sample of the frame as the relative time reference . Thus, occurrences ahead of the reference have a negative value. The values of OFE and OLE are treated as the features of the mid-5 ms segment of the frame.

It is seen in Fig.3.2(a) that though there is a change in the signal structure of the voiced frame, the entire signal within the frame belongs to a homogeneous class as seen by the nearly uniform sinusoid in the bandpass filtered signal.

For both the voiced and unvoiced speech segments shown in Figs. 3.2(a) and 3.2(b), OFE and OLE occur long before and after the reference instant i.e. $OFE \ll 0$ and $OLE \gg 0$. Figure 3.2(c) shows a transition from a voiced to an unvoiced segment. Here, OFE is highly negative and OLE has a low negative value. The converse is true for an unvoiced to voiced transition as shown in Fig. 3.2(d).

From the above illustrations, we can deduce the following: (a) When $OFE \ll 0$ and $OLE \gg 0$, the frame corresponds to a homogeneous class. (b) $OFE \ll 0$ and $OLE \approx 0$ for a frame with a transition from a high to a low amplitude (H-L). (c) $OFE \approx 0$ and $OLE \gg 0$ for a frame with a transition from a low to high amplitude (L-H). Thus, we can divide the speech signal into a homogeneous class (H-class or L-class) and the two types of transitions H-L and L-H.

3.2.3.3 Mean absolute difference between extrema (MADE) in a frame

The peak values of the BPF signal in Fig.3.2(b) are very low (≈0.005). The transitions in such frames are ignored, since the whole frame corresponds to a non-sonorant segment. For this purpose, another

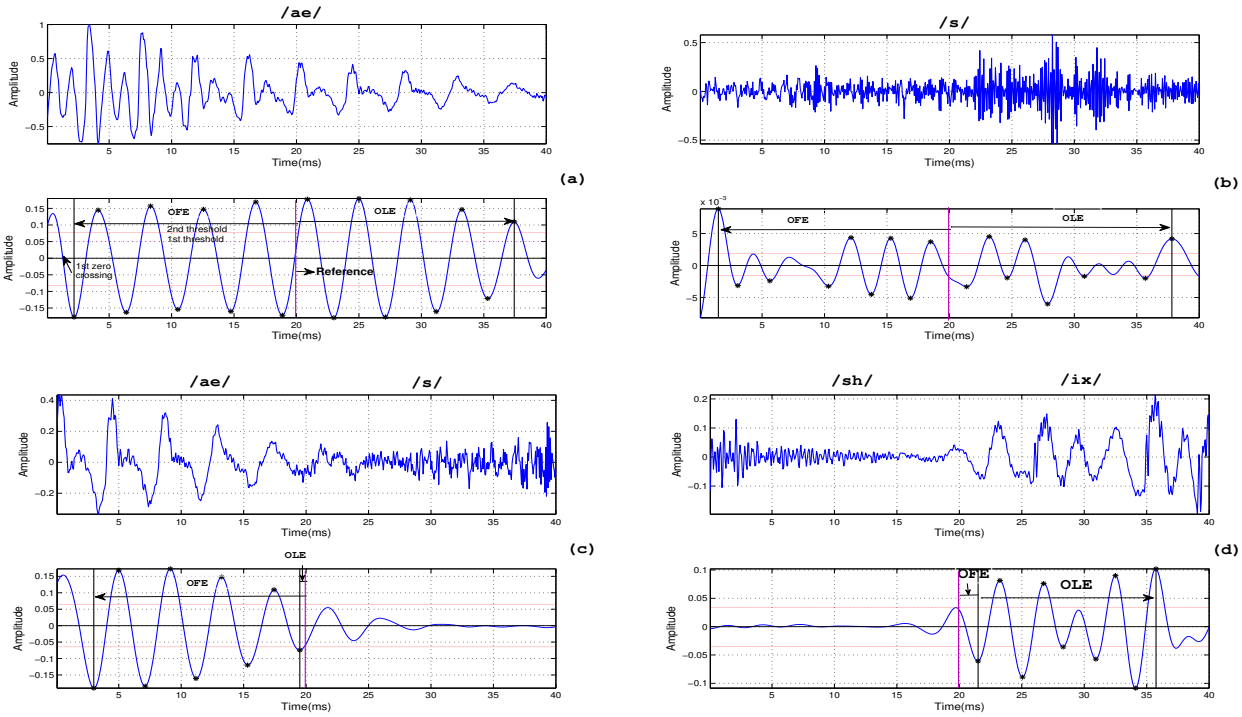


Figure 3.2: Speech signal (top) of some sample frames and their corresponding bandpass filtered versions (bottom). The extrema above the second-pass thresholds (horizontal lines above and below zero) as well as the occurrences of the first (OFE) and last extrema (OLE) are shown. (a) A homogeneous voiced segment (MADE= 0.32). (b) A homogeneous unvoiced segment (MADE= 0.01). (c) A voiced-unvoiced transition (MADE= 0.31). (d) An unvoiced-voiced transition (MADE= 0.27).

measure named mean absolute difference of extrema (MADE) is introduced, which is the mean of the absolute differences between successive peaks and valleys after the second thresholds. The caption for Fig.3.2 also gives the values of MADE for each of the sample signals shown. Figure 3.3 shows the histogram of MADE for sonorant and non-sonorant frames for twenty randomly selected files from the TIMIT database, used as a development set. The histogram suggests an optimal threshold of 0.024 for sonorant/non-sonorant classification. Transitions corresponding to OFE and OLE are ignored when MADE is below this threshold. Spurious detection of transitions in unvoiced segments and due to frication noise following bursts are avoided and we detect only transitions from/to sonorant segments.

3.3 Algorithm for the detection of transitions

We refer to the proposed algorithm shown in Fig. 3.5 as AGR algorithm. The rules cited in the flowchart are presented in Tables 3.1-3.5. The rules have been derived by studying the nature of variation of the proposed features (OFE, OLE, MADE and SI) as a function of the change of analysis window from frame to frame for different types of transitions (silence to non-silence, sonorant to unvoiced non-sonorant and vice versa). The study was carried out on a small development set of the TIMIT [43] speech database. The thresholds and parameter values involving time durations defined

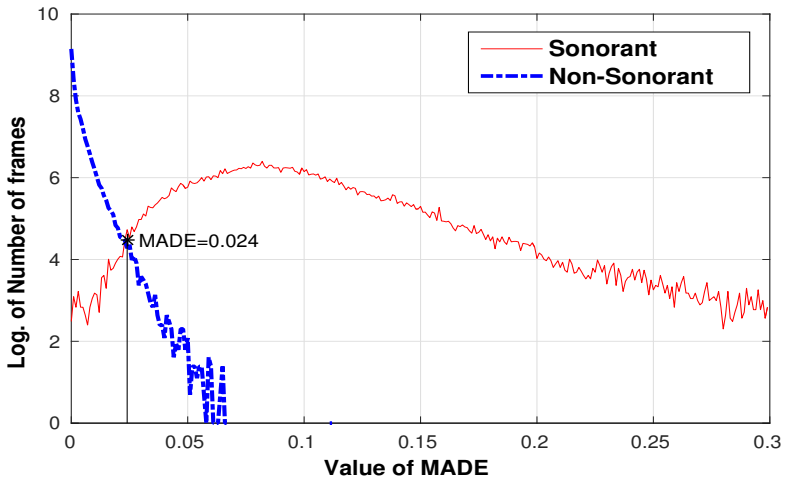


Figure 3.3: Histogram showing the distribution of computed values of frame-wise mean absolute difference between extrema (MADE) for 20 randomly selected files from TIMIT database.

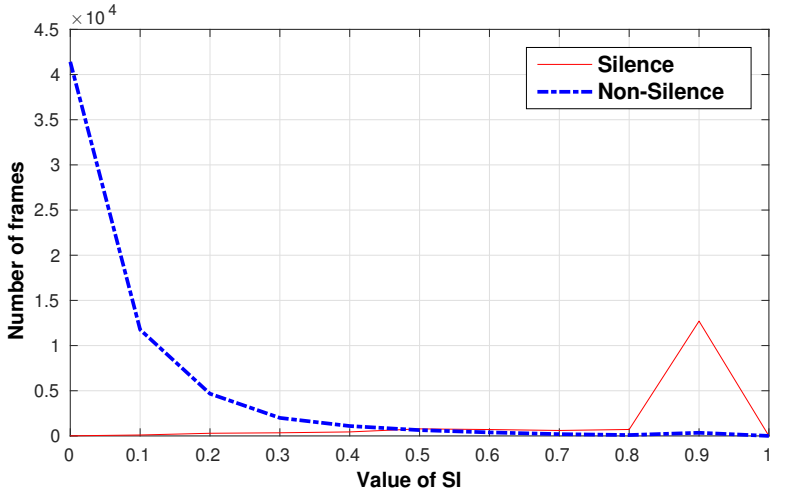


Figure 3.4: Histogram of frame-wise silence index of 20 randomly selected files from TIMIT database.

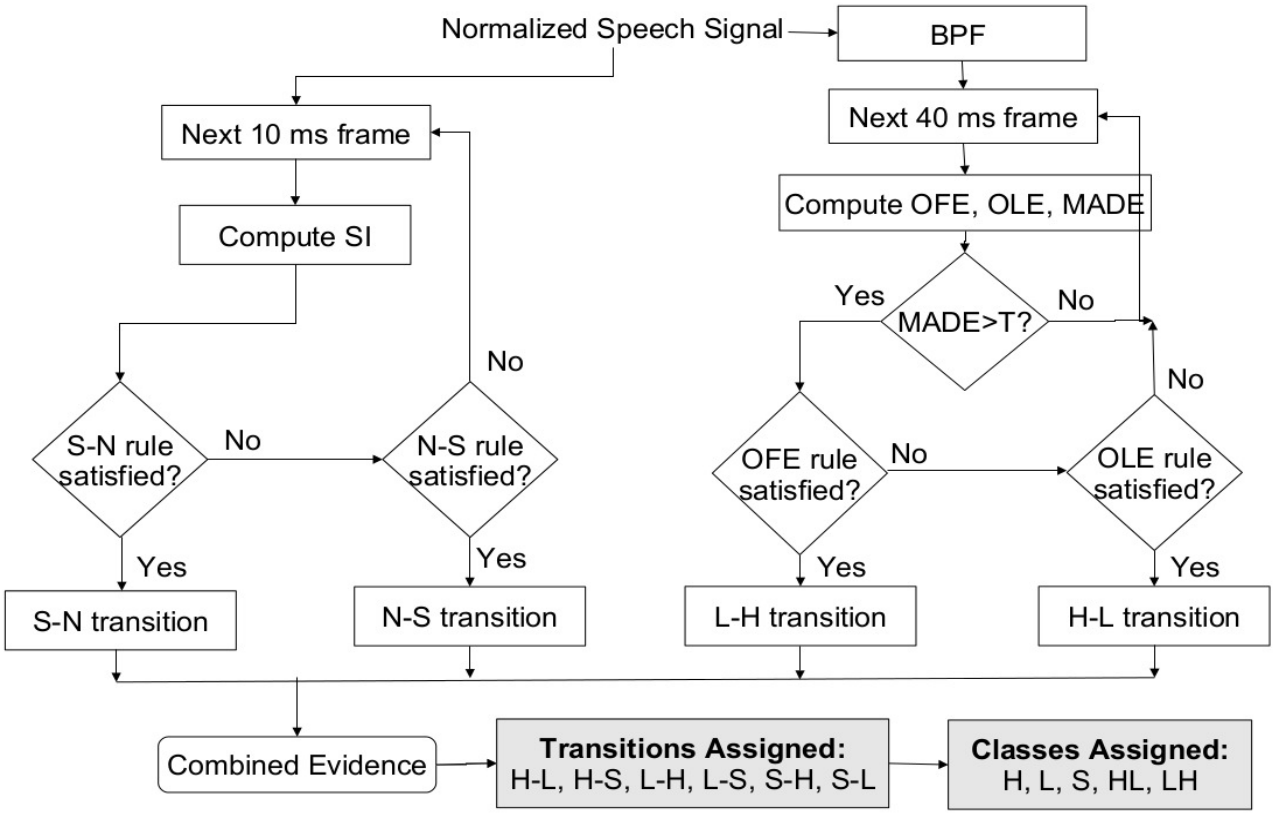


Figure 3.5: Flowchart for the detection and class assignment of transitions (T is the threshold for MADE).

in the rules have been assigned based on the statistics of the duration of phonemes marked in the TIMIT training database.

We discuss the strategy used in the algorithm with an example. The first step divides the speech signal into silence and non-silence segments.

3.3.1 Detection of transitions between silence and non-silence classes

To arrive at a threshold, the histogram of SI values for silence (S) and non-silence (N) frames for the development set is computed and is plotted in Fig.3.4. A threshold value of $SI = 0.5$ is chosen to distinguish between silence and non-silence frames. The characteristic changes in the value of SI for three consecutive frames are used to detect the transition from silence (S) to non-silence (N) classes and vice-versa. These rules are listed in Table 3.1. The temporal accuracy for the transition is improved by recomputing SI in the mid 5 ms region for non-overlapping sub-segments of 1 ms duration. The instant of transition is defined as the point when SI crosses the value of 0.5 in any direction. The samples between an N-S and a following S-N transitions are labeled as S-class and vice versa, as shown in Table 3.3.

Table 3.1: Rules for detecting transitions between silence (S) and non-silence (N) classes.

SI values and the past context	Type
$SI(k-1) \geq 0.6$ and $SI(k+1) \leq 0.4$	S-N
$SI(k) \leq 0.35$ and *Previous transition* =N-S	S-N
$SI(k-1) \leq 0.4$ and $SI(k+1) \geq 0.6$	N-S
$SI(k) \geq 0.7$ and *Previous transition* =S-N	N-S

SI values obtained for a speech segment containing several phones and the detected transitions are shown in Fig. 3.8(a). The S-N transitions around 200 ms corresponding to the boundary between 'h#' and /sh/ and around 770 ms corresponding to the boundary between /dcl/ and burst /d/ are detected successfully. However, the /dcl/ segment (550-580 ms) between /eh/ and /jh/ is missed, since considerable energy is present in the corresponding segment. The phone /jh/, normally a voiced affricate, is realized as unvoiced in this utterance. It is not clear if a closure needs to necessarily be marked for an affricate realized as a fricative. Despite the presence of a noticeable impulse, /kcl/ segment from 920 to 980 ms is correctly identified as S-class.

3.3.2 Detection of transitions between sonorant and non-sonorant classes

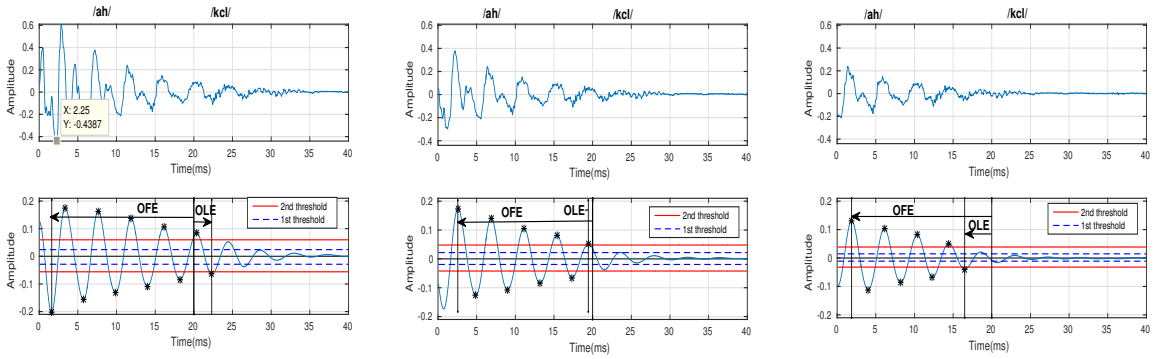
A transition from a non-sonorant to a sonorant class is mostly detected as L-H and vice versa.

Suppose there is a vowel, followed by an unvoiced or voiced stop. Figure 3.6 shows three consecutive frames of a speech segment where a strong H-L transition occurs from a vowel /ah/ to a unvoiced closure /kcl/. It is seen that OLE decreases from a positive value (Fig.3.6(a)), crosses zero (Fig.3.6(b)), and then further decreases to a negative value (Fig.3.6(c)). Figure 3.7 shows three consecutive frames of a speech segment where a weak H-L transition occurs from a vowel /ih/ to a voiced closure /dcl/. It is seen that OLE decreases from a positive value (Fig.3.7(a)) to a minimum near zero (Fig.3.7(b)), and then without any zero crossing again increases (Fig.3.7(c)).

As long as the analysis window contains only the vowel part, most peaks and valleys in the BPF signal have comparable, high amplitudes. Hence, the first and the last extrema occur at the beginning and end of the analysis frame. This makes the values of OFE and OLE highly negative and positive, respectively, with respect to the centre of the frame. Once the closure region enters the analysis window, the occurrence of the last extremum (still from the vowel region only) slowly moves towards the centre, reducing the OLE value from a high positive value towards zero (See Figs. 3.6 and 3.7, (a) and (b)). However, OFE remains highly negative, since there is a part of the vowel still at the beginning of the analysis window. Now, since nearly half the analysis window contains the closure signal, the first-pass threshold reduces.

In the case of unvoiced stop, there are no peaks in the closure interval and hence, the second-pass threshold remains almost the same, being decided only by the extrema of the vowel. Thus, after the next frame shift, OLE moves further to the left, beyond the centre (See Fig.3.6(c)). Thus, OLE reduces from a high positive value, becomes zero and then goes negative. Thus, OFE having a consistently high negative value and OLE having a zero crossing from positive to negative value denotes a high

(low-frequency) amplitude phone (say, a vowel) to a low amplitude phone (say, unvoiced stop or fricative) transition or a strong H-L transition. Similarly, OLE having a high positive value and OFE having a zero crossing from positive to negative value denotes a low to high amplitude transition.

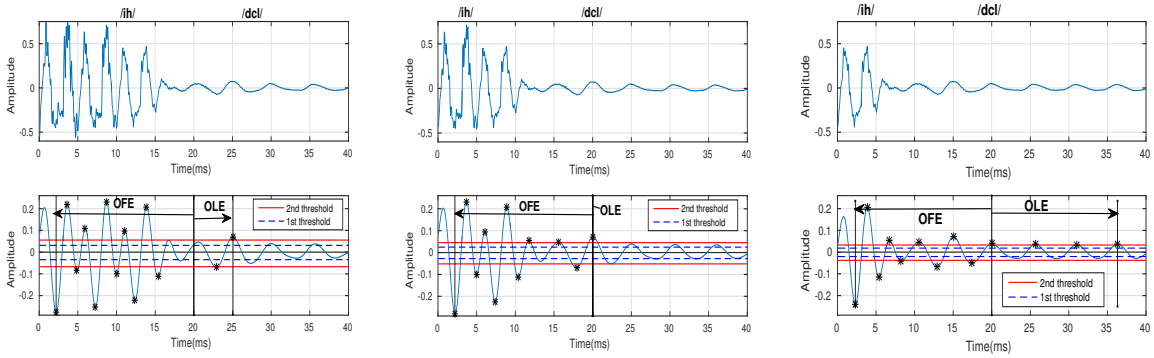


(a) (b) (c)

Figure 3.6: The signals and their bandpass filtered versions of three consecutive frames of speech containing a voiced (/ah/) to a unvoiced closure (/kcl/) transition (strong H-L). The occurrences of the first (OFE) and last extrema (OLE), and the first and second-pass thresholds are shown, in each case. Plots of OFE and OLE contours show that OLE goes through a positive to negative zero crossing.

In the case of voiced stop, there are small, but definite peaks in the closure interval, which bring down the second-pass threshold also. After the next frame shift, the voiced closure region enters the first half of the analysis window, further bringing down the second-pass threshold (See Fig.3.7(c)). Now, most of the peaks and valleys in the closure region survive the low second-pass threshold . Thus, OLE again becomes highly positive, rather than becoming negative. Thus, OFE having a consistently negative value and OLE going through a minimum within 5 ms (the duration of a frame shift) from the centre also denotes a H-L transition. However, to distinguish it from the above scenario, we call this as a weak H-L transition. Thus, in a weak H-L transition, the OLE, rather than going through a positive to negative zero crossing, actually goes through a minimum and again increases. Similarly, OLE having a high positive value and OFE going through a maximum near zero and again decreasing denotes a weak L-H transition. Figure 3.7 shows three consecutive frames of a speech segment where a weak H-L transition occurs from a vowel /ih/ to a voiced closure /dcl/. It is seen that OLE decreases from a positive value (Fig.3.7(a)) to a minimum near zero (Fig.3.7(b)), and then without any zero crossing again increases (Fig.3.7(c)).

Table 3.2 shows the rules for detecting L-H and H-L transitions. As the frame window shifts from a non-sonorant to a sonorant segment, OFE changes from negative to positive, and reduces towards zero. So, a L-H transition is detected near the positive to negative zero crossing of OFE as seen in Fig.3.2(d). Appropriate class labels are assigned to the segments between successive transitions. For example, in the simplest case, a segment which lies between H-L and L-H transitions would naturally be labeled as L-class. The labelings used are shown in Table 3.3.



(a) (b) (c)

Figure 3.7: The signals and their bandpass filtered versions of three consecutive frames of speech containing a voiced (/ih/) to a voiced closure (/dcl/) transition (weak H-L). The occurrences of the first (OFE) and last extrema (OLE), and the first and second-pass thresholds are shown, in each case. Plots of OFE and OLE contours show that OLE goes through a minimum near zero.

Table 3.2: Rules for detecting transitions based on contours of OFE and OLE. NZC, MIN and MAX denote a positive to negative zero crossing, a local minimum and maximum, respectively of OFE or OLE contour.

Nature of OFE and OLE in the frame	Type
NZC in OFE, $OLE \gg 0$	Strong L-H
OFE is a MAX, $\|OFE\| \leq 5\,ms$, $OLE \gg 0$	Weak L-H
$OFE \ll 0$, NZC in OLE	Strong H-L
$OFE \ll 0$, OLE is a MIN, $\|OLE\| \leq 5\,ms$,	Weak H-L

Figure 3.8(b) shows the bandpass version of the segment of speech signal shown in Fig.3.8(a). The values of OFE and OLE obtained for successive frames are scaled and plotted as a function of time in Fig. 3.8(b). Towards the end of /sh/, just before 300 ms, OFE rapidly increases from a negative to a positive value and returns to a negative value for the next phone. The positive to negative zero crossing (NZC) in OFE marks a strong L-H transition. We choose the zero-crossing in the BPF signal closest to this NZC as the transition instant.

Table 3.3: Class labeling of a segment between successive transitions.

Types of successive transitions		Class label assigned
k^{th}	$(k+1)^{th}$	
N-S	S-N	S
S-N	N-S	N
L-H	H-L	H
H-L	L-H	L
H-L	H-L	LH
L-H	L-H	HL

During the H-L transition from /ih/ to /dcl/ around 720 ms, there is an abrupt decrease in

amplitude (unlike /eh/ to /dcl/). OLE decreases rapidly to a minimum, close to the base line, without a sign change. If this minimum value of OLE is within 5 ms, then it is considered a weak H-L transition (fourth row of Table 3.2). The value of 5 ms arises because of the frame shift. In the next analysis frame, the voiced closure enters the first half of the window, thus reducing the second-pass threshold. This renders the extrema of the voiced closure region to go above the threshold, thus taking the value of OLE back to a high positive value. A similar weak L-H transition (second row of Table 3.2) due to OFE is seen between /dcl d/ and /ah/ around 800 ms. Thus the so called weak transitions are also as genuine transitions. The distinction between a strong and weak transition is noted only for the sake of further analysis, if required.

It is not necessary that L and H classes always alternate. It may be noted that across /kcl-k/ and /s-ux/, there are two consecutive L-H transitions due to OFE. In order to distinguish such transitions, the segment between two consecutive L-H transitions is denoted as HL class. Similarly the signal between two consecutive H-L transitions is labeled as LH class. Occurrences of LH and HL classes are rare. However, this specific example of HL class is an exception. Though the segment /k s/ should have been labeled HL class, the first transition across /kcl-k/ due to OFE is ignored since MADE is below threshold and hence the label happens to be L. The transition /kcl k/ is still captured as S-N transition based on SI as shown in Fig.3.8(a). Thus, the class label of a speech segment need to be decided by combining the information provided by SI and OFE/OLE.

3.3.3 Class assignment based on combined evidence

SI is computed with a frame duration of 10 ms on the speech signal whereas the frame duration used for the measurement of OFE and OLE is 40 ms, obtained from the BPF signal. Thus, OFE and OLE are computed independently of SI. Since the transitions between H and L are detected independent of the transitions between N and S, these two evidences are combined to get a single stream of transitions. The decisions assigned by the two processes correspond to the same 5 ms inter-frame segment, as both the frames of 10 ms and 40 ms duration have the same center. The very first frame of any utterance is assumed to be a silence frame. After detecting the transitions, their locations are arranged in an ascending order of occurrences in time. It is highly likely that a detected transition due to OFE or OLE may lie close to a transition detected using SI. For example, a vowel offset followed by a stop closure would be detected as a H-L and a N-S transition; a silence followed by a sonorant gives rise to both L-H and S-N transition.. Such simultaneous transitions are merged into a single transition. Hence, decisions need to be made on the temporal spacing allowed between the two types of transitions to merge them into one and the same transition and on the location of the new, merged transition.

The combined evidences for merging the two types of transitions are listed in Table 3.4 while Table 3.5 lists the rules for assigning class labels to the segments lying between successive transitions. The following points are noted with respect to combining evidences:

(a) Any H-L or L-H transition detected within a silence class is marked for removal and the type of transition is noted for N-class assignment. This obvious condition is not shown in the Table.

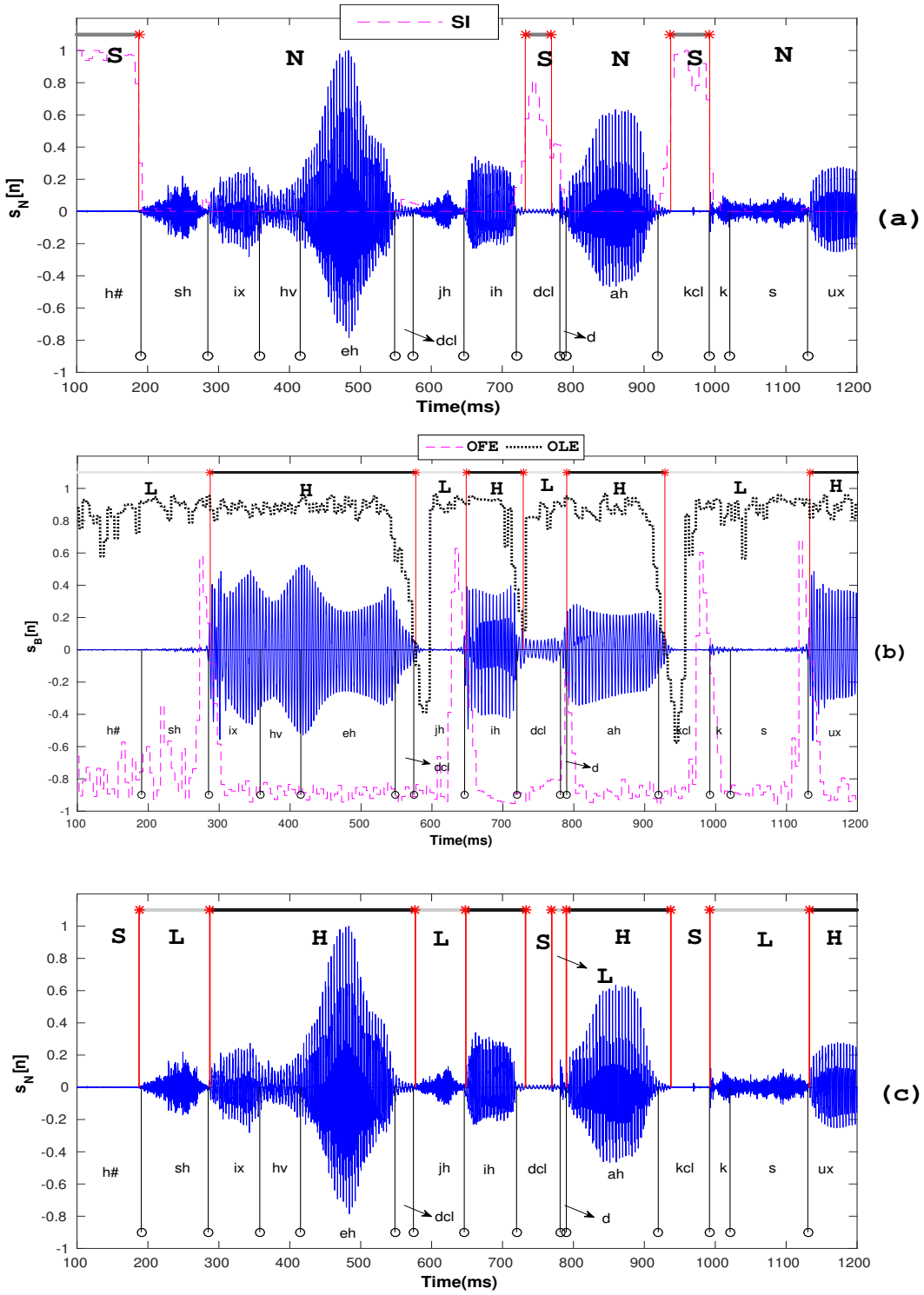


Figure 3.8: (a) S-N and N-S transitions (starred markers) detected using SI values derived from the original signal. (b) L-H and H-L transitions detected using the OFE/OLE values derived from the BPF signal. (c) The merged transitions along with the original signal.

(b) When a S-N transition is followed by a L-H transition within 10 ms or by a H-L transition within 20 ms, then the latter transition $(k+1)^{th}$ is marked for removal. As a consequence, the $(k+2)^{th}$ transition will now become $(k+1)^{th}$ transition.

(c) When a H-L or L-H transition is followed by a N-S transition within 20 ms, then the former transition is marked for removal.

Table 3.4: Combining evidences for merging adjacent transition labels.

Successive Transitions		Separation (ms)	Outcome
k^{th}	$(k+1)^{th}$		
S-N	L-H	$\leq$ 10 ms	Remove $(k+1)^{th}$
S-N	H-L	$\leq$ 20 ms	Remove $(k+1)^{th}$
L-H	N-S	$\leq$ 20 ms	Remove k^{th}
H-L	N-S	$\leq$ 20 ms	Remove k^{th}

The segment between a S-N and the next transition is labeled as L or H-class, respectively, depending upon whether the next transition is of type L-H or H-L. If the next transition is of type N-S (because of a H-L or L-H transition that was removed), then either of the class labels H or L which occupied most of the duration between S-N and N-S is assigned to the segment (not shown in Table 3.5). Further, the segment before a N-S transition is labeled as H- or L-class if the preceding transition is a L-H or H-L transition, respectively. If the preceding transition is of type S-N (because of a removal), it is handled similar to the situation discussed above.

Thus, the following five classes result after combining the evidences from the two types of transitions: (a) H (b) L (c) S (d) HL (f) LH.

Figure 3.8(c) shows the class labels assigned after the evidence combination. A S-N transition around 200 ms is followed by a L-H transition around 280 ms. Since the two transitions are spaced beyond 10 ms, both are retained. The N-class segment between S-N and L-H is assigned the L-class. The segment (approximately between 280 and 580 ms) consisting of /ix hv eh dcl/ homogeneously belongs to H-class, since it lies between L-H and H-L transitions. Similarly, /ih/ (650 to 720 ms) and /ah/ (790 to 920 ms) belong to H-class. The segment /jh/ (580 to 650 ms) belongs to L-class.

Some more situations are discussed, where evidence combinations are applicable. The boundary (around 730 ms) between /ih/ and /dcl/ is detected both as N-S and H-L transitions. However, since

Table 3.5: Class assignment for the segment between k^{th} and $(k+1)^{th}$ transitions after combining evidences. k and $k+1$ denote the revised frame indices.

Types of successive Transitions		Class assigned
k^{th}	$(k+1)^{th}$	
S-N	L-H	L
S-N	H-L	H
L-H	N-S	H
H-L	N-S	L

H-L transition was on the left of the N-S transition and within 20 ms (see Fig. 3.8(a)), it is removed. Although the H-L transition is removed, the segment /ih/ between the L-H and N-S transitions is labeled as H-class.

A S-N transition is detected across /dcl/ and /d/, and a L-H, across /d/ and /ah/. These two transitions, being 20 ms apart are retained as the temporal tolerance is 10 ms for a L-H following a S-N transition. Across /ah/ and /kcl/, both H-L and N-S are detected, in that order. However, since H-L transition is within 20 ms of N-S transition, the former is removed by the combined evidence. Across /kcl-k/, only the S-N transition survives since L-H transition due to OFE is removed due to the low value of MADE.

3.4 Experimental details and evaluation

The proposed AGR algorithm has been validated on the entire TIMIT database [43], i.e., both training and test databases for clean speech. It consists of several dialects of North American English, totaling 6300 utterances spoken by 630 speakers. For evaluation on noisy speech, only the test database is used, as listed in Table 3.6. The TIMIT database has hand labels at the phone level and the closure duration of stops have been explicitly marked. Accordingly, the class 'stops' denotes 'stop bursts'.

Table 3.6: Databases used for evaluation of performance on clean and noisy speech

	Clean speech	Noisy speech
Database used	TIMIT training and test set	TIMIT test set : 168 speakers, 1344 sentences

Every detected transition is uniquely assigned to the nearest TIMIT boundary. The statistics of the temporal differences between the labeled boundaries and the assigned transition instants are computed. The boundary detection accuracy is measured for different values of temporal tolerance.

In order to study the relationship between the manner of articulation and the homogeneous segments, the distribution of each class of phones among the five classes assigned after combined evidence step is computed. Phonetic grouping given in TIMIT database as shown in Table 1.1 is used as the reference for assigning the class of phones.

For every sonorant and non-sonorant onset in the labelled database, we verify if there is a detected transition within a specified temporal tolerance. If no transition has been detected within the tolerance for an onset, then it is a case of miss or deletion. This measures the accuracy of detection of onsets relative to the type of transition. A detected transition for which there is no associated labeled boundary is counted as an insertion. The ratio of the number of insertions to the total number of transitions detected is another performance measure.

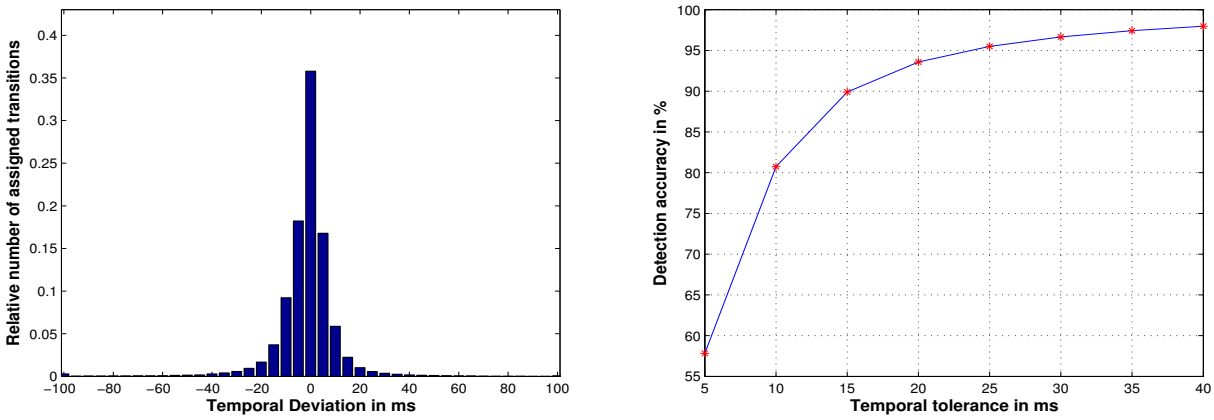
3.5 Results and discussion

We first present results on clean speech. The results presented correspond to the total number of frames of 3,818,197 and the total number of detected transitions of 144,715.

3.5.1 Temporal accuracy of detection

Figure 3.9(a) shows the histogram of the temporal deviations of the detected transitions from the hand labeled boundaries, using a bin size of 5 ms. The mean and standard deviation are -1.62 and 17.05 ms, respectively. We observe that 36.4% of the detections are within $\pm$2.5 ms.

The detection accuracy is computed for different values of the temporal tolerance, namely, 5 to 40 ms in steps of 5 ms. The ratio of successful detections to the total number of transitions, excluding insertions, is computed as the detection accuracy of transitions and is shown in Fig. 3.9(b) as a function of temporal tolerance. 57.8% of the transitions lie within $\pm$ 5 ms and 98% of the transitions lie within $\pm$ 40 ms. Thus, the temporal resolution of detection is higher than those of related previous works as presented in Sec.3.5.5.



(a) (b)

Figure 3.9: (a) Histogram of the temporal deviations of the detected transitions from the hand labeled boundaries, (b) Detection accuracy in percentage (%) of transitions as a function of temporal tolerance.

3.5.2 Classes of phones detected across each type of transition

It is of interest to know the distribution of various classes of phones (vowels, semivowels etc.) which belong to the five broad classes obtained: H, L, S, HL and LH. This distribution is listed in Table 3.7, for a temporal tolerance of 20 ms. More than 91% of vowels belong to H-class. But we note that there are about 4.8% of vowels in L class and 0.5% in S class. About 41% of the 'ax-h' phones (not shown in the table) lie in L-class, since it has the characteristics of unvoiced speech, with a very low amplitude in the BPF signal. Amongst the semi-vowels, 74.0% of 'hh' lie in the L-class, since this phone also has characteristics similar to unvoiced speech. It is seen that in any nasal-fricative segment, there is a short interval of silence at the end of the nasal, which is however not hand labeled as silence. This explains the occurrence of about 6.8% of nasals in S-class. A short silence segment is not unexpected since there is a change of source process as well as a drastic shift in the articulatory positions.

About 91% of affricates and unvoiced fricatives lie in L-class. Affricates include the voiced affricate

Table 3.7: Relative distribution of each class of phones among the broad five classes. Results on the entire TIMIT data, containing both training and test data.

Segment type	H	L	S	HL	LH
Vowels	91.8	4.8	0.5	2.2	0.7
Semivowels	86.2	9.2	1.3	2.5	0.9
Nasals	82.0	7.2	6.8	3.3	0.8
Unvoiced fricatives	4.2	91.1	2.0	2.2	0.4
Voiced fricatives	28.6	56.9	8.3	5.0	1.2
Voiced stops	48.0	37.3	12.1	1.6	1.1
Unvoiced stops	16.3	70.5	11.3	1.3	0.5
Affricates	6.3	91.0	0.1	1.8	0.7
Others	4.5	13.3	82.1	0.1	0.0
Voiced closures	10.6	8.7	78.4	1.7	0.7
Unvoiced closures	2.1	5.6	92.0	0.2	0.1

/jh/, which also lies in L-class. Amongst the fricatives, 20% of 'th' lies in S-class, since it sometimes manifests as a burst with a closure interval.

56.9% of voiced fricatives also lie in L-class, whereas 28.6% lie in H-class and 8.3% go to S-class. The presence of voicing in voiced stops gives rise to a large amplitude BPF signal and when these classes follow a silence or a L-class phone, they go to H-class. 72% of /z/ lies in L-class despite being a voiced fricative. The phone 'dh' sometimes behaves like a stop with a closure and /v/ is realized both as voiced and unvoiced.

The phone labels of the TIMIT database are mapped to the phonetic classes, sonorants and non-sonorants. All vowels, semi-vowels and nasals are assigned to the sonorant class. Others ('h#', 'epi', 'pau') and the closures of stops are assigned to the silence class. Non-sonorants include all the phones except the sonorants and silence. The relative distribution as per the phonetic classes, 'sonorant', 'non-sonorant' and 'silence' is shown in Table 3.8. About 89.8% of sonorants lie in H-class. About 75.5% of non-sonorants are in L-class. If we remove voiced fricatives and voiced stops from non-sonorants, then the unvoiced non-sonorants in L-class increases to 84%. This suggests that we need two groups of non-sonorants. 84.2% of 'silence' segments lie in S-class with 10.5% in L-class. Once again, this may arise due to some so called silence phones like 'h#' and 'epi' having a high amplitude.

Based on the above results, we can broadly state that H-class represents the sonorant class and L-class represents unvoiced non-sonorants, whereas voiced non-sononrants may be found either in H or L classes.

3.5.3 Onset of sonorants and non-sonorants vis-a-vis the type of transition

The onsets of sonorants and non-sonorants are considered as landmarks [121], [132],[133]. It would be of interest to relate the onsets of sonorants and non-sonorants to the detected types of transition. We have excluded /q/ from non-sonorants as done in several previous works [130],[132]. Further, within the non-sonorants, we consider the fricatives and stop bursts separately to detect the onsets. The

Table 3.8: Distribution of each broad class of phones in the TIMIT database among the five classes.

Phone class	H	L	S	HL	LH
Sonorant	89.8	5.8	1.3	2.4	0.7
Non-sonorant	15.2	75.5	6.2	2.4	0.7
Silence	4.8	10.5	84.2	0.4	0.1
Voiced non-sonorant	34.8	50.6	9.5	3.9	1.2
Unvoiced non-sonorant	8.4	84.2	5.0	1.9	0.5

Table 3.9: Percentage of onsets of broad phonetic classes detected as a function of temporal tolerance.

Onset of	Type	20	30	40
Sonorants+	L-H, S-H	92.0	94.0	94.7
Unvoiced fricatives/affricates*	H-L, S-L	83.0	85.4	86.5
Stop closures	L-S, H-S	77.2	80.0	81.4
Bursts	S-H, S-L	87.7	88.7	89.1

+Following an unvoiced fricative, unvoiced stop or an affricate. *Following a sonorant or a silence

results are shown in Table 3.9. For a tolerance of 30 ms, 94% of onsets of sonorants occur at L-H or S-H transitions. We have considered sonorants following unvoiced fricatives, unvoiced stops and affricates, since voiced fricatives and voiced stops may lie in H class (see Table 3.7). The onsets of unvoiced fricatives and affricates occur at H-L and S-L transitions 85.4% of the time within 30 ms. Stop closures are detected as onsets 80% of the time across L-S and H-S transitions. Onsets of stop bursts invariably (88.7%) follow a detected silence segment (S-H, S-L). The results are comparable even for a tolerance of 20 ms. Hence the proposed method also serves the purpose of landmark detection with a good accuracy and temporal resolution.

3.5.4 Insertions

The insertions on the whole TIMIT database are 8.7%. Table 3.10 shows the relative distribution of insertions among the various class of phones.

Table 3.10: Relative distribution of insertions amongst classes of phones (V-voiced, UV-unvoiced)

Class	Vowels	Semivowels	Nasals	UV fricatives	V fricatives	V stops	UV stops	affricates	others	V closures	UV closures
Insertions(%)	14.63	2.87	4.76	13.03	5.70	2.84	21.11	0.41	12.68	9.52	12.46

About a third of these insertions occur during the silence, i.e, 'others' and closures of stops. Segments like 'h#' and 'epi' may contain impulse like noise with significant amplitude resulting in some spurious S-N and N-S transitions. About 24% of the insertions occur during stops. These arise partly due to multiple bursts. A transition is also detected across a low level aspiration interval following a strong burst. While this is a desirable feature of the algorithm, since the aspiration interval is not explicitly marked, such transitions get reported as insertions. During unvoiced fricatives, especially, /f/, the amplitude of the signal varies considerably with intermittent low frequency, large amplitude pulses resulting in a high rate of insertions (about 11%).

3.5.5 Comparison with the previous work

In terms of detecting classes, this work is comparable to manner classification [120],[99] and in terms of detecting onsets, this work is closest to the landmark detection reported in the literature [132],[133].

The present work differs from the previous related works in four important aspects: (a) The temporal features used in this study are different from those proposed in the earlier studies. (b) The proposed algorithm has been tested on the entire TIMIT database, whereas the previous studies have reported results based on a limited test data (16 speakers speaking a total of 80 utterances for the development set and 16 new speakers speaking 48 utterances for the test set taken from the TIMIT database in a study by Liu [133]; 504 utterances from the test set of the TIMIT database in the study by Salomon et al [132]). (c) The transitions or landmarks to be detected correspond to different events. Liu [133] and Salomon et al [132] defined three landmarks. (d) The quoted results correspond to a temporal tolerance of 30 ms [133] or 50 ms [132]. Due to the above disparities, we can only make a broad qualitative comparison with the previous works. The comparison of our results with the published results of Liu [133] and Salomon et al. [132] is summarized in Table 3.11.

Salomon et al. [132] tested their method on the manner classes of sonorant, fricative, stop and silence. The average accuracy using 39 dimension MFCCs or 12 parameters derived from four temporal features was reported as 70% for a tolerance of 50 ms, whereas with the combined features, it increased to 74.8%. Compared to these results, the accuracies of the proposed method are 89.8%, 84.2% and 84.2% for sonorants, unvoiced non-sonorants and silence classes, respectively, within 20 ms tolerance, when tested on the entire TIMIT database (see Table 3.8).

In Liu's [133] study, of the total number of landmarks, 83% and 88% were within 20 and 30 ms of the labeled boundaries, respectively. The classes considered in that study are sonorants, fricatives and bursts. The above results may be compared with the temporal accuracy of detection of the present work (Sec.3.5.1). For a temporal tolerance of 20 and 30 ms, our temporal accuracy is 93.6% and 96.7%, respectively.

Table 3.11: Performance comparison of various algorithms with respect to temporal accuracy of detection.

Method	Database	Results
AGR (our method)	Whole TIMIT training and test set	93.6%, 96.7% within 20 ms, 30 ms
Liu [133]	48 utterances from TIMIT database	83%, 88% within 20 ms, 30 ms
Salomon [132]	504 utterances from test set of TIMIT database	74.8% within 50 ms

Since about 91% of vowels belong to H-class and about 94% of L-H, S-H transitions within 30 ms represent the onset of a sonorant segment, we can compare these results to the vowel onset point (VOP) detection [123]. In some respect, the detection of onsets of sonorants may be compared with the onset detection of vowels [123],[124]. For tests conducted on isolated utterances of CV units with 5220 VOP events, Prasanna et al. [123] report a detection accuracy of 88% within a tolerance of 30 ms. In the case of continuous speech, for the small chosen set of 25 sentences having a total of 236

VOP events, they report a detection accuracy of 88.5% within a tolerance of 20 ms. For some selected utterances from the TIMIT database, for 173 tokens, for seven speakers, Prasanna et al. [124] report an accuracy of 67.6% for 30 ms tolerance. Compared to these results, for the entire TIMIT database, we get a detection accuracy of 91% for sonorant onsets for 20 ms temporal tolerance.

3.6 Robustness in the presence of noise

We evaluate our algorithm on noisy speech generated by adding different kinds of noise to the test set of TIMIT database at various signal to noise ratios (SNR's). The following noises are used for evaluating our algorithm for detection of transitions:

- Schroeder noise [135]: It is a localized white noise. As the energy level in a speech utterance varies widely with time, the clean speech is corrupted with Schroeder noise so that samplewise SNR is constant in the noisy speech. We use the model as devised in [135], where the noisy speech signal is generated by the formula $y[n] = s[n](1 + \epsilon\eta[n])$, where $s[n]$ is the speech signal, ϵ is the factor determining the noise energy which changes with the desired SNR and $\eta[n]$ is the randomly chosen $+1$ or -1 with equal probability.
- White noise: It is generated from a zero mean normal distribution, with the standard deviation being determined by the SNR desired.
- Babble noise: It is taken from the Noisex-92 database [18] and scaled appropriately to generate noisy speech with the desired SNR.

Figure 3.10 shows the percentage of the total number of transitions (with respect to 38,198 transitions detected in the case of clean speech) detected by our algorithm for the three types of noisy speech with SNR varying from 0 to 30 dB. It is seen that at a low SNR of 10 dB, our algorithm detects 8.62% for speech with babble noise, 39.59% in the case of white noise and 96.40% in the case of Schroeder noise.

Insertions at an SNR of 10 dB are 0.23% for white, 0.49% for babble and 9.36% for Schroeder noise. Since the number of transitions detected is low for white and babble noise, it is imperative that the % number of insertions is less. It is observed that transitions between silence and non-silence segments (S-N and N-S) are missed for white and babble noises at low SNR's since the silence regions are corrupted by noise, leading to poor performance on detection of transitions. In the case of Schroeder noise, as the samplewise SNR is constant, silence segments are not corrupted with high noise and the corresponding transitions are preserved, detecting 33,026 (86.46%) transitions even at 0 dB SNR.

Figure 3.11 shows the precision of detection within a temporal tolerance of 20 ms for the three noises as a function of SNR. It is seen that among the detected transitions, even at an SNR of 5 dB, precision above 91% is achieved. It is seen that temporal accuracy does not change much with variation in SNR for Schroeder noise, since the low energy in the silence segments is preserved due to uniform local SNR and hence the S-N and N-S transitions remain intact. For white and babble noise,

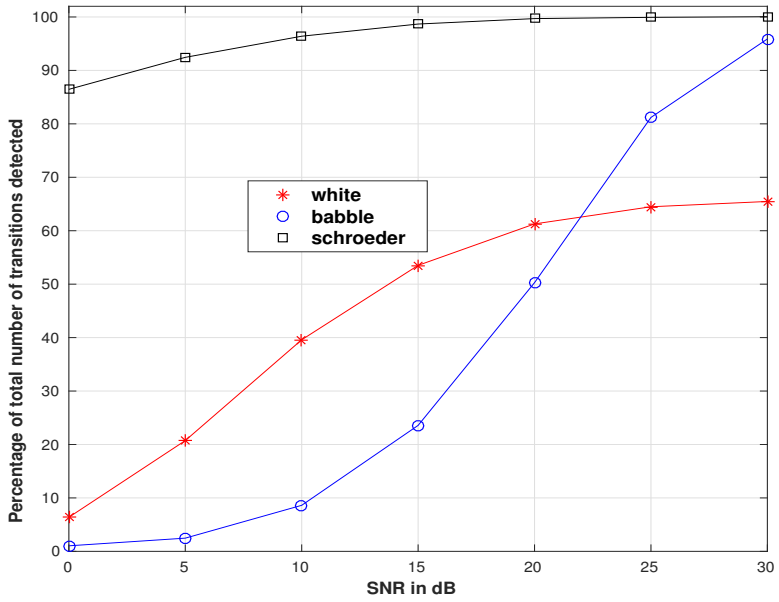


Figure 3.10: Percentage of total number of transitions detected on TIMIT test set as a function of input SNR

energy in the silence segments increases with increase in noise energy (or decrease in SNR) and hence the SI value is low even for silence segments at low SNR, which leads to missing S-N, N-S transitions.

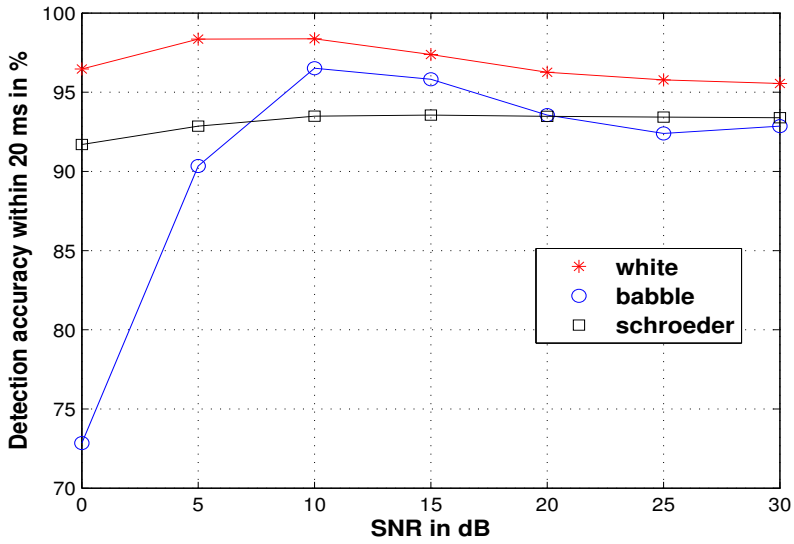
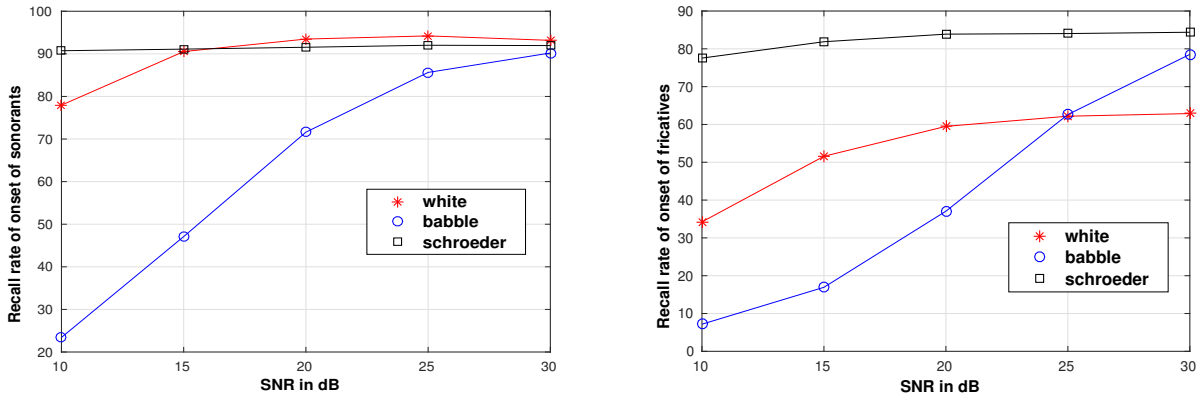


Figure 3.11: Precision of detected transitions (for a temporal tolerance of 20 ms) as a function of SNR

Figure 3.12 shows the percentage of onsets of sonorants and fricatives detected (recall) within a tolerance of 20 ms. Since babble noise has significant low frequency energy, accuracy of detection of

onsets of sonorants and fricatives suffers at low SNR's. For white and babble noises, even though we miss S-N and N-S transitions, H-L and L-H transitions are preserved at low SNR's, which result in relatively high detection of onsets of sonorants and fricatives.



(a) (b)

Figure 3.12: Percentage of onsets of (a) sonorants and (b) fricatives detected (recall) within a tolerance of 20 ms as a function of SNR in dB for white, babble and Schroeder noises

3.7 Conclusion

For the DFs and PFs to be complimentary to the statistical approach, we believe that an acoustic-phonetics knowledge based approach needs to be pursued. In our understanding, the highlights of such an approach is that it does not require a huge amount of training data and a small development set is considered sufficient. In this chapter, we have proposed a knowledge-based approach to the problem of detecting transitions in both clean and noisy speech signal. Further, several studies have pointed out the robustness of temporal features in speech perception [132],[136]. In the proposed method, using only four simple measures, we have been able to demonstrate that landmarks like the onsets of sonorants (L-H, S-H), unvoiced sonorants (H-L, S-L), closures of stops and stop bursts can be detected with a high accuracy ($> 85\%$) and with a good temporal resolution (20 ms). These results are as good or better than state-of-the-art methods which make use of high dimensional acoustic features and sophisticated classifiers. Although a number of techniques exist for segmentation, alternate approaches are to be explored, since they may complement one another and offer robustness.

3.7.1 Future Work

During the course of this investigation, we have made some observations, which are noted here for future work: (a) We could inquire how OFE/OLE measures perform instead of the abrupt energy change measures used in the literature for the detection of landmarks [133], manner classes [132], bursts [111] and vowel onset points [124]. (b) Our preliminary investigation shows that OFE and OLE measures computed on a speech signal, instead of bandpass signal, are useful to identify certain

transitions within vocalic segments. Also, OFE and OLE may be computed on subband signals. (c) The number of extrema in a speech signal relative to the number of extrema in the corresponding bandpass signal is a useful parameter for distinguishing between voiced and unvoiced segments. (d) We have observed that bursts most often lie at the end of a silence or L-class. This narrows down the search interval for detecting the bursts. These preliminary observations need to be formalized and tested in a future work.

Chapter 4

Estimation of GCIs using subband analysis of linear prediction residual of speech

Utilizing the presence of rich harmonics in the quasi-periodic excitation of the vocal tract, we explore the possibility of extracting the glottal closure instants (GCIs) from one or more subbands of the linear prediction (LP) residual of the pre-emphasized speech signal. We also study the noise robustness of this approach. A composite signal is derived as the sum of the smoothed envelopes or amplitude modulation components of the subband signals. GCIs obtained from this signal are refined using the LP residual signal, and are validated against the GCIs obtained from the electroglottograph signal. We propose three subband methods, each differing in either the filterbank used for decomposition or the envelope extraction scheme. Six different databases are used for evaluating the performance of the proposed methods. The effectiveness of the algorithms is studied on telephone, reverberant and noisy speech for different signal to noise ratios down to - 5 dB, with different kinds of additive noise, namely white, pink, blue, babble, vehicle and HF channel noises. The performance is also compared with three of the state of the art algorithms. The results show that the performance of the proposed methods is consistently comparable to the best of the other techniques for both clean and noisy speech. The HBEBEST study shows the possible potential of subband methods to achieve very good performance in the presence of additive noise.

4.1 Introduction

In the previous chapter, we detected transitions across various classes and segmented the speech signal into different classes. The transitions were detected across abrupt change of source process (voiced/unvoiced/silence). In this chapter, we further look at the instants where specific changes occur in the glottal source quasi-periodically. So, in contrast to the previous chapter, we analyze the speech signal at a deeper level and further segment it into quasi-periodic cycles.

Glottal closure instants (GCIs) or epochs are the instants at which the vocal tract is maximally excited when the glottis closes abruptly. Hence, GCI corresponds to one of the peaks in the LP residual signal. These instants of significant excitation in speech are important events used in many speech analysis and synthesis algorithms. Some of the applications of GCIs are pitch synchronous pitch and duration modification [137], speaking rate modification, pitch normalization, speech coding/compression, and speaker normalization.

It was first shown in [138] that the LP residual contains GCI information, where they have estimated GCIs as the significant maxima of the hilbert envelope of the filtered LP residual. In [139], GCIs are estimated as the positive zero-crossings of the phase slope function of the LP residual. This is improved by the Dynamic Programming Phase Slope Algorithm (DYPSA) [140], where dynamic programming is employed to correct the baseline phase slope based pitch marking algorithm by minimizing the pitch deviation and the phase slope costs.

In zero frequency filtering (ZFF) method [141], GCIs are estimated as the positive zero crossings of the signal obtained by filtering the speech signal around a single frequency and mean subtraction around a window. In Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) method [142], the search for GCI is narrowed down to a short interval starting at the minimum of a windowed mean based signal, and then picking local maxima from the LP residual within the interval. Both of these algorithms require a priori average pitch period information for assigning the window length.

The wavelet approach [143] assumes that the speech signal has predominant peaks and dynamically tracks the movement of the valley in each wavelet band to arrive at the GCI. Sub-band analysis of speech to find pitch frequency is proposed in [144] using the auditory models of speech perception, and AM-FM approach to find the pitch is proposed in [145]. In [146], GCIs are estimated from the amplitude modulation (AM) component of the Bessel approximation of the speech signal in the 0 to 300 Hz range .

The Yet Another GCI Algorithm (YAGA) is an LP-based approach which estimates GCIs from a set of GCI candidates by finding the best path using dynamic programming [147]; it differs from DYPSA in the way the candidate set is estimated. In [148], epochs are estimated using a non-linear temporal measure called dynamic plosion index applied on the processed integrated linear prediction residual (ILPR) of the speech signal.

In [149], six popular algorithms are reviewed for evaluating GCIs on speech spoken with modal voice and seven additional voice qualities. In [150], an exhaustive evaluation of five state of the art GCI detection algorithms is presented using six different databases. In this chapter, we evaluate and compare the proposed algorithms with three of those methods, namely DYPSA, ZFF and SEDREAMS, on the same six databases.

Figure 4.1 shows the narrowband spectrograms of a segment of voiced speech and its corresponding LP residual. It is evident from the dark horizontal stripes in Fig.4.1 that the LP residual contains harmonics of the pitch frequency in voiced segments. Thus, the GCI information, which is related

to the pitch, must also be present in the harmonics. We therefore propose to utilize this redundancy to achieve robustness through subband analysis. An advantage of processing LP residual over speech signal is that in speech, a few subbands are affected by formants which distort the harmonics and the GCI information as seen in Fig.4.1. Also, additive noise affects only certain frequency bands of speech. Hence, reliable GCI information may be present in unaffected frequency bands.

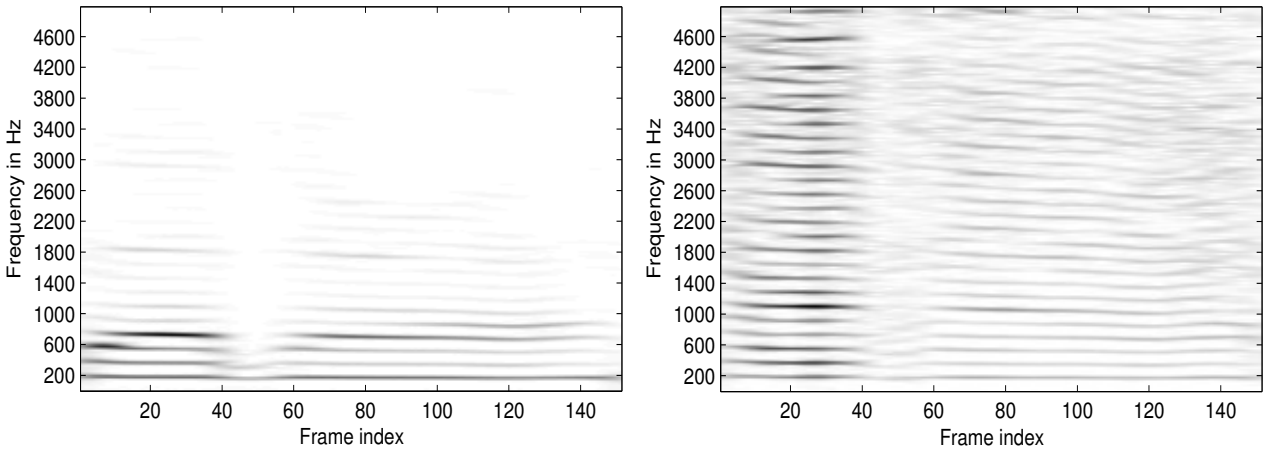


Figure 4.1: Spectrograms of (a) a segment of voiced speech, and (b) its LP residual

Based on this motivation, we have obtained the GCIs from the subband decomposition of the LP residual of the pre-emphasized speech signal within a specified bandwidth. We have proposed another approach for GCI estimation in [151], where the subbands of the speech signal itself have been used. In [152], product of the multiscale estimations of wavelet transform of speech signal have been used for determining GCIs while in [153], local maxima of the dyadic wavelet transform of the speech signal at different scales are combined to detect the transients caused by the glottal closure.

In general, it is easier to estimate the GCIs for clean speech, whereas accurate estimation of GCIs in noisy speech is challenging. Accordingly, in our work, we derive a composite signal (CS) from the individual subbands, which is able to reasonably preserve the nature, namely high amplitude and high rate of change, of the instants of significant excitation. We estimate the initial glottal closure instants from the CS, which are refined using the peaks in the LP residual.

The contributions of the work reported in this chapter are: (a) A novel subband approach to GCI estimation; (b) Applying it to three different subband decomposition approaches and (c) Experimental evidence to show the efficacy of the proposed methods in the presence of various types of additive colored and real world noises at different SNR's.

Section 4.2 explains the proposed approach for estimating GCIs. Section 4.3 gives the details of the simulation. Simulation results along with performance comparison with other methods are presented in Sec.4.4. Discussion and conclusions about the proposed approach are presented in Sec.4.5 in the light of the simulation results.

4.2 Proposed Approach

The proposed approach is summarized in Fig. 4.2.

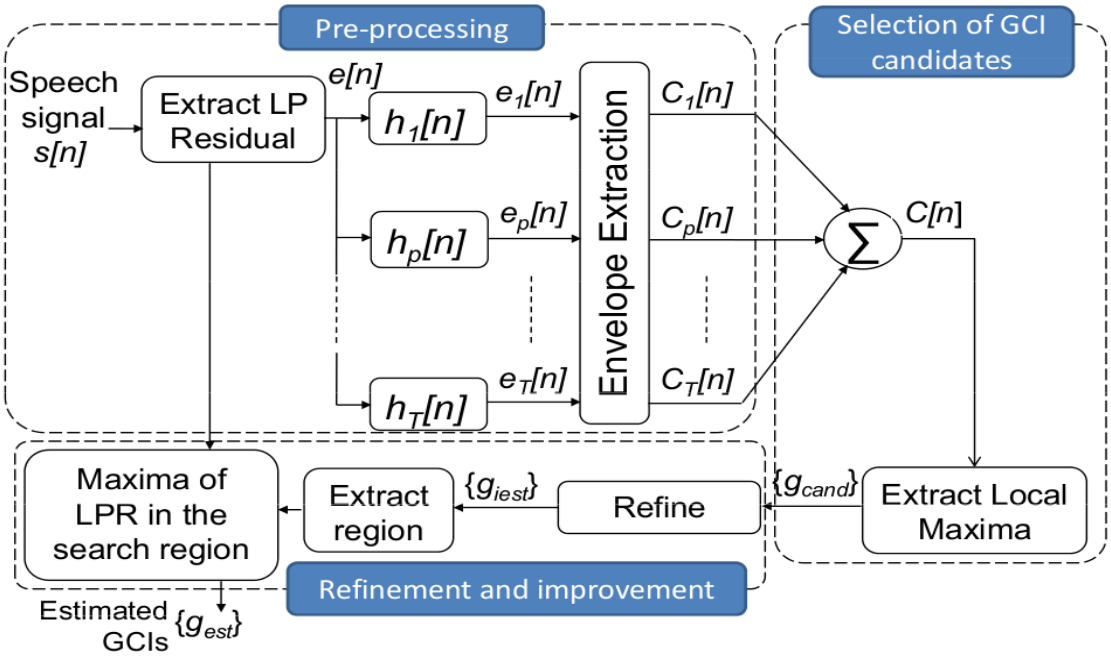


Figure 4.2: Overview of the algorithm for GCI estimation.

A speech signal, $s[n]$ may be expressed as $s[n] = e[n] * v[n]$, where $e[n]$ is the excitation signal and $v[n]$ is the vocal tract impulse response. The excitation signal, $e[n]$ can be estimated by inverse filtering the pre-emphasized speech signal using the LP coefficients, which is the LP residual (LPR) signal. $e[n]$ is passed through a filterbank to obtain the subband signals. Let the impulse response of the p^{th} filter in the filterbank shown in Fig. 4.2 be denoted by $h_p[n]$. The p^{th} subband of the LPR is the output of this filter:

$$e_p[n] = e[n] * h_p[n] \tag{4.1}$$

where $*$ denotes convolution. The details of the nature of filters to be used in the filterbank is presented in Sec.4.3.1. Through extensive experimentation, we have observed that the peaks of the envelope of the absolute value of each subband signal correspond to the candidate GCIs. We therefore resort to envelope extraction. The envelope of the p^{th} subband signal is referred to as the p^{th} subband

component, and denoted by $C_p[n]$. Figure 4.3 shows the absolute value of the subband signal, $|e_5[n]|$ obtained using a hamming filterbank and its envelope, $C_5[n]$ obtained by cubic hermite interpolation of its local maxima.

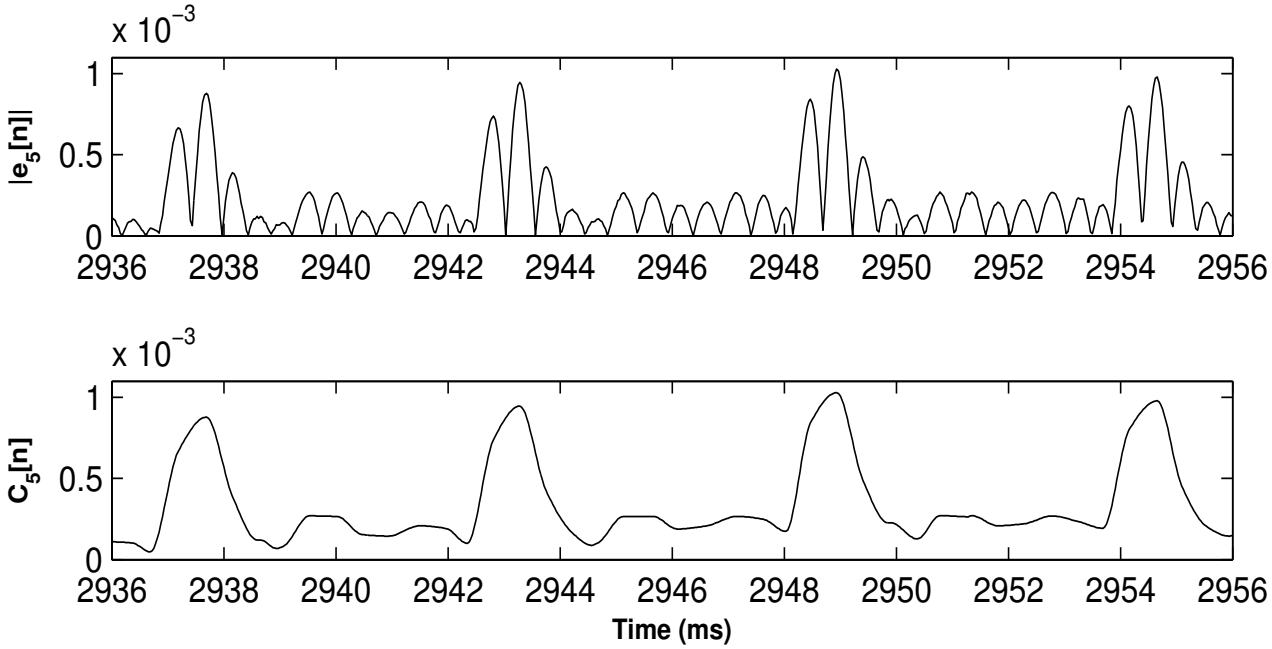


Figure 4.3: Absolute value of subband signal, $|e_5[n]|$ and its envelope,$C_5[n]$

4.2.1 Composite signal from the subband components

Since subband components are quasi-periodic, their aligned local maxima can be considered as candidates for GCI estimation. Due to the filtering and the envelope extraction processes, it is observed that the accuracy of GCI estimation may be compromised by using candidates from individual subband components. Also, based on the varying pitch frequency and its harmonics across various frames of the speech signal, GCI information will be present in different subband components. We therefore utilize the redundancy in the subbands and derive a composite signal as the sum of subband components. The composite signal, being the average of multiple subband components, provides more robust candidates for GCI estimation and compensates for the varying pitch frequency. It is also observed that no weighting of subband components is required for computing the composite signal since the LP residual signal has a relatively constant power spectral density. In the presence of additive noise, since different subbands are affected differently by noise, we believe that reliable GCI information will still be present in a few subbands where the signal to noise ratio (SNR) is high.

The composite signal, $C[n]$ is obtained as

$$C[n] = \sum_{p=1}^{T} C_p[n] \tag{4.2}$$

where T is the total number of bands into which the signal is decomposed in the subband envelope approach.

4.2.2 Estimating initial GCIs from the composite signal

The composite envelope signal $C[n]$ preserves the characteristics of the impulsive excitation instants required for GCI estimation. It is analyzed frame-wise with a frame size of 50 ms and a frame shift of 20 ms. The time instants of the local maxima between successive zero crossings in the mean subtracted composite signal are the potential candidates for GCIs, g_{cand} as shown in Fig. 4.4. In the refinement block in Fig. 4.2, estimates of a reference pitch period, $p_r[i]$ and amplitude, $a_r[i]$ are obtained as :

$$p_r[i] = \sum_{j=-10}^{1} (g[i+j+1] - g[i+j])/12 \tag{4.3}$$

$$a_r[i] = (C[g[i-1]] + C[g[i+1]])/2 \tag{4.4}$$

where $g[i]$ is the present GCI candidate and $g[i+j]$ is the initial estimate of past GCI, g_{iest} (for $j < 0$), or the future GCI candidate, g_{cand} (for $j \geq 1$).

The candidate GCIs (g_{cand}) are refined (see Fig. 4.2) to obtain the initial estimates of GCIs (g_{iest}) by applying certain constraints on local periodicity and relative amplitudes. The thresholds and conditions are defined by experimentation on a small development set and apriori knowledge that GCIs occur quasi-periodically. The present GCI candidate, $g[i]$ is pruned (confirmed as false detection) in the refinement block, if any one or more of the following conditions are satisfied:

1. $C[g[i]] < C[g[i-1]]$ *and* $(g[i] - g[i-1]) < 2/3(p_r[i])$
2. $(g[i] - g[i-1]) < 0.25 p_r[i]$
3. $C[g[i]] < 0.1 a_r[i]$

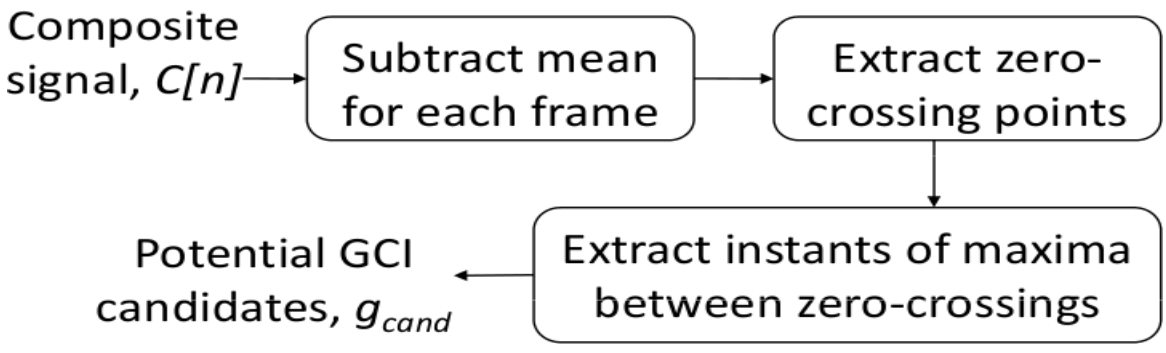


Figure 4.4: Extraction of GCI candidates from the composite signal.

The effectiveness of our approach is shown to be valid in Fig.4.5. Here, the excitation signal is simulated by a train of pulses, representing a series of GCIs. Speech-like signal is simulated by

filtering this by a forward LP filter (vocal tract). The LP parameters are derived from a real speech signal. Now, we add white noise at SNR of 0 dB and implement our subband envelope method on this simulated noisy speech. We observe that the GCIs estimated by our method are in proximity to the excitation pulses with no spurious detections.

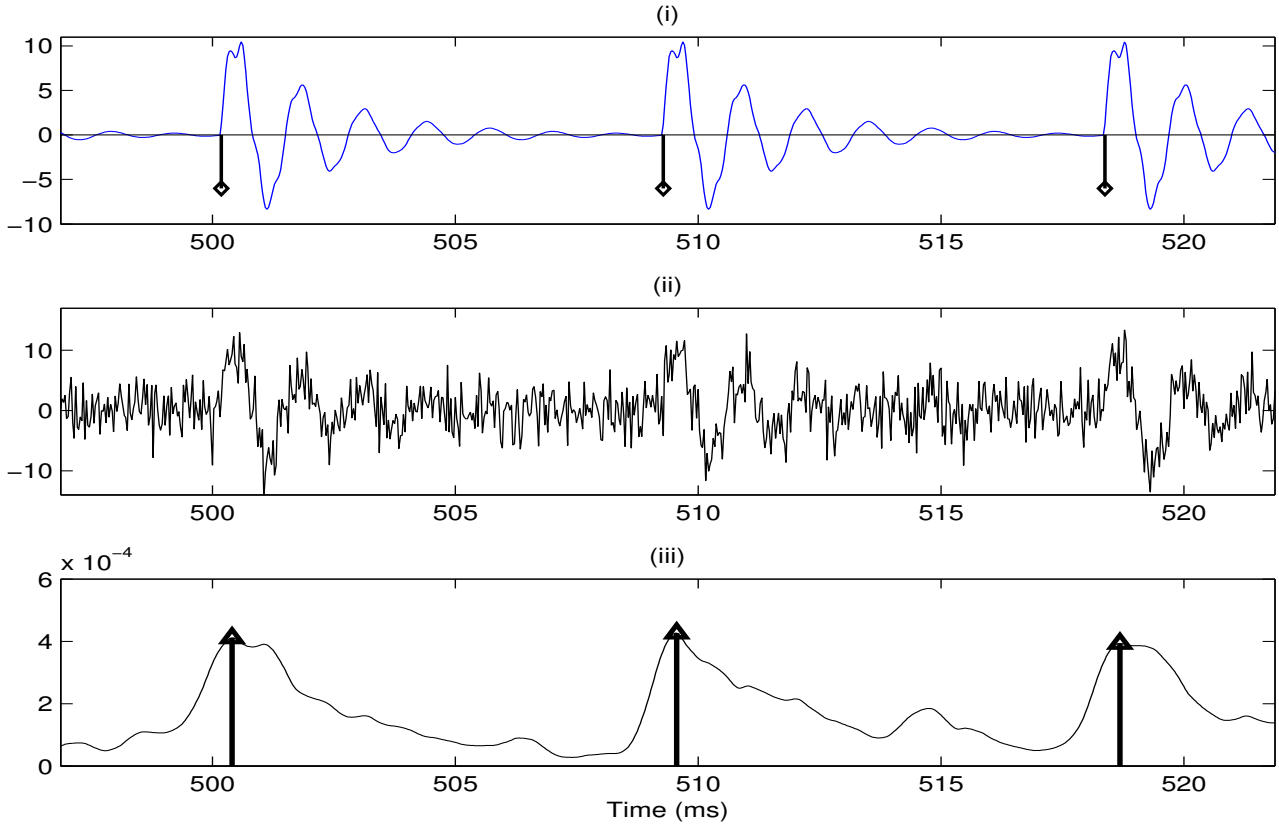


Figure 4.5: Illustration of the basis of the proposed approach, using simulated speech. (i) *solid curve* is the simulated speech; *diamonds* denote the instants of the excitation pulses. (ii) simulated speech with additive white noise at SNR of 0 dB. (iii) *solid curve* is the composite envelope of subbands; *triangles* indicate the GCIs estimated by subband envelope method.

4.2.3 Estimating final GCIs from the LP residual signal

As the composite signal is the average of the subband components, it is seen in some cases that accurate detection of the position of GCIs may not occur due to smearing of the peaks in $C[n]$. It is known that the GCIs are concentrated near the significant local maxima of the LP residual signal. Hence, we use the LP residual of the speech signal to refine the initial estimate of GCIs, g_{iest}, as shown in Fig. 4.2.

A region centered around the initial estimate, g_{iest}, is extracted with a width 0.15 times the reference pitch period. The final estimated GCI, g_{est}, is obtained as the maxima of the LP residual signal within this region.

4.3 Simulation Details

The details of the proposed approach based on the subband envelopes, databases used and the performance measures are given in this section. Analysis is carried out only for the voiced regions of the signal, which are identified from the derivative of the electroglottograph (dEGG) signal [154]. A segment between two consecutive dEGG minima is considered voiced if their absolute values are above 10% of the absolute value of the global minimum, and if the pitch frequency is in the range of 70 to 400 Hz.

LP coefficients are obtained from Hanning windowed frames of length 20 ms taken from the pre-emphasized speech signal using an LP order of f_s (sampling frequency) in kHz+2. The residual signal is obtained by inverse filtering the pre-emphasized signal using the LP coefficients.

4.3.1 Filterbank design

The filterbanks are designed such that the bandwidth of each filter encompasses approximately one harmonic of the pitch period and the center frequency of the highest filter is upto 1800 Hz. The choice of 1800 Hz as the highest center frequency is motivated by our experimental observation on a small development set that higher frequency components do not contribute to the accuracy of GCI estimation. In what follows, we propose three different methods for obtaining the subband components, $C_p[n]$:

1. Hamming filterbank followed by envelope extraction, which we call as Hamming Bandpass Envelope (HBE) method. The envelope of the p^{th} subband signal is obtained by piecewise cubic hermite interpolation [155] between the local maxima of $|e_p[n]|$.

   The Hamming filterbank used is type I, symmetric, odd length, bandpass filterbank, whose impulse response for each band p is given by

$$h_p[n] = \begin{cases} \left(\dfrac{\sin\left(2\pi f_2[p]\left(n - D/2\right)\right)}{\pi\left(n - D/2\right)} - \dfrac{\sin\left(2\pi f_1[p]\left(n - D/2\right)\right)}{\pi\left(n - D/2\right)} \right) w[n], \\ 2(f_2[p] - f_1[p]), \text{if } n = D/2. \end{cases}$$

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{D}\right) \tag{4.5}$$

   where D is the order of the filter, $0 \leq n \leq D$, $f_1[p]$ and $f_2[p]$ are the cutoff frequencies and $w[n]$ is the hamming window function. $h_p[n]$ is symmetric about $D/2$ and the frequency response of this filterbank is shown in Fig. 4.6a. We choose the center frequencies of the filters to be separated by 200 Hz, thereby yielding a set of 9 filters. The first center frequency is 100 Hz. The order of each filter is chosen to be $D = 64$. The above values of the center frequency and the number of filters have been arrived at by experimentation using a small development set. It

is observed that varying the bandwidth and center frequencies do not change the performance of GCI estimation.

It may be intuitive to choose the center frequencies as the harmonics of the average pitch frequency of the speech utterance. So, we experimented with a variation of the HBE method by choosing the center frequencies for the 9 filters as multiples of the average pitch frequency. We call this variant of HBE as HBEVAR (HBE with variable center frequency) and observed that we get results similar to HBE. The disadvantage of using HBEVAR is that we need an apriori estimate of the average pitch frequency.

2. Hamming filterbank followed by AM extraction, which we call as Hamming subband AM-FM decomposition (HAM) method. Here, the envelope is obtained by extracting the AM component. AM-FM decomposition of R subband signals extracted from the LP residual signal $e[n]$ is seen as [156]:

$$e[n] = \sum_{k=1}^{R} e_k[n] \tag{4.6}$$

$$e_k[n] = a_k[n]\cos[\omega_k n + \phi_k[n]] \tag{4.7}$$

Here, each $e_k[n]$ represents a subband signal [156], for which $a_k[n]$ is the AM component, ω_k is the resonant frequency, and $\phi_k[n]$ represents the phase/frequency modulation component. In the AM-FM perspective of speech, the vocal tract is excited by puffs of air modulated by the glottis. This signal is modulated by the vocal tract to produce speech. Here, we may consider the source excitation signal as AM signal and vocal tract as the frequency modulator. Hence, at the instants of maximum excitation, the AM component is expected to have a peak.

Even though the LP residual signal is an approximation of the excitation signal, it may still have some vocal tract component. Computing the AM components and resonant frequencies from the LP residual signal is an ill-posed problem. The problem is circumvented by splitting the signal into subbands using a hamming filterbank such that each band has at most one dominant AM-FM component. We consider the AM components to estimate GCIs, and extract them for the k^{th} subband using the Discrete-time Energy Separation Algorithm-1 (DESA-1) [157]:

$$\psi(e_k[n]) = e_k^2[n] - e_k[n-1]e_k[n+1]; \tag{4.8}$$

$$a_k[n] = \sqrt{\frac{\psi[e_k[n]]}{1 - \left(1 - \frac{\psi[e_k[n] - e_k[n-1]]}{2\psi[e_k[n]]}\right)^2}} \tag{4.9}$$

where ψ is called the Teager energy operator (TEO) [158]. The absolute AM component $|a_k[n]|$ is smoothed using a symmetric exponential filter to get the subband component $C_k[n]$ in Fig. 4.2,

since the original AM component has many spurious local maxima, which cause false detection of GCIs.

In the case of HAM, the center frequencies of the hamming filters are separated by 125 Hz, thus yielding a set of 14 filters. The first center frequency is 62.5 Hz.

3. Gammatone filterbank followed by envelope extraction, which we call as Gammatone subband Envelope (GTE) method.

   We have used an equal bandwidth, equally spaced gammatone (GT) filterbank, which is different from the equivalent rectangular bandwidth (ERB) scale normally used for modeling auditory filters. ERB scale has increasing bandwidth and spacing, which we found to be not suitable for GCI estimation, since higher frequency bands encompass multiple harmonics. Experimental validation also confirmed that equally spaced GT filterbank were better than ERB scale for GCI estimation. Gammatone filter is a linear filter described by an impulse response that is the product of a gamma distribution and a sinusoidal tone. The impulse response of a GT filter for each band p is given by,

$$h_p[n] = a(n/f_s)^{L-1} e^{-2\pi Bn/f_s} \cos[2\pi f_p n/f_s + \phi] \tag{4.10}$$

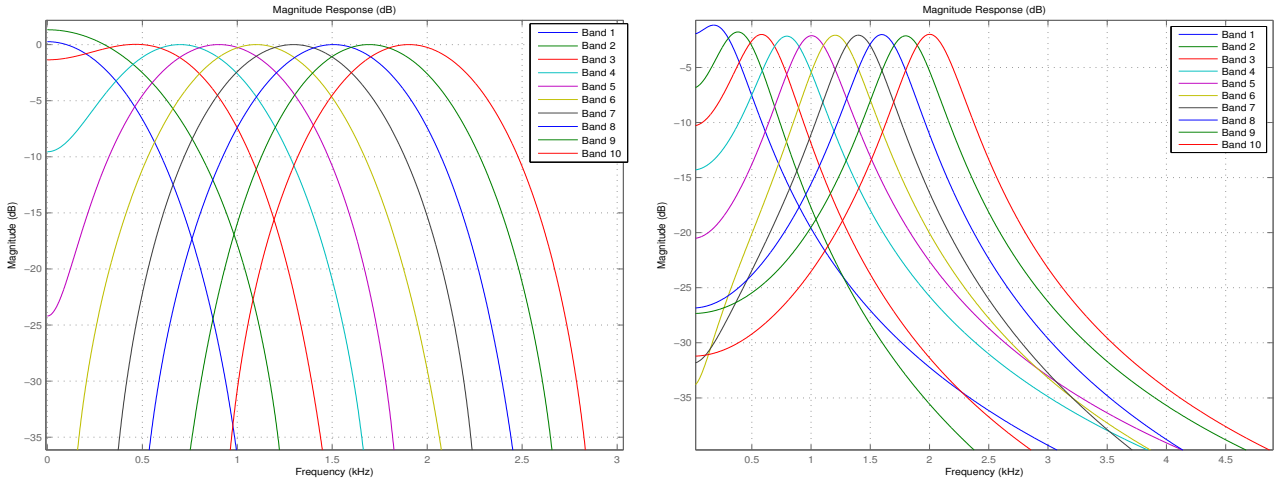
   where a is the amplitude, L is the order of the filter, f_p is the center frequency for the p^{th} subband, B is the filter's bandwidth, f_s is the sampling frequency and ϕ is the phase of the carrier. The frequency response of the GT filterbank is shown in Fig. 4.6b. The order of the filters is chosen to be 2, with the center frequencies of adjacent filters separated by 200 Hz, yielding a set of 9 filters. The first center frequency is 200 Hz. The envelope is obtained similar to the HBE method.

The phase delay of the hamming filter used is constant across all frequency components since it is a linear phase filter. Hence by using adequate delay compensation, one can get accurate estimation of GCIs. However, the phase delays of gammatone filters are different for different bands, being equal to the position of the peak of the individual impulse response as given in [159]. These delays are compensated to align the peaks in the envelopes of the GT subband signals.

4.3.2 Databases, noises and ground truth for GCIs

We have used six databases [150] for our study, all of them containing the EGG signal along with the stereo speech signal:

- The first three databases are taken from CMU arctic databases [160] (two male and one female speakers). Utterances have a duration of 1 to 6 seconds each, adding up to about 56 minutes for each database. The speech signals have been sampled at 32 kHz and quantized at 16 bits.
- The fourth database contains non-sense words spoken by RAB, a UK male speaker, containing all the English phone-to-phone transitions.



(a) Hamming filterbank

(b) Gammatone filterbank

Figure 4.6: Magnitude response of the filterbanks used as a function of frequency.

- The fifth one is KED TIMIT database (male speaker).
- The sixth database, APLAWD, consists of five sentences, each of which has been uttered ten times by five male and five female speakers. The last three databases have been sampled at 16 kHz.

To test the robustness of our subband based methods, we applied our techniques on simulated telephone bandwidth speech (passband between 300 to 3400 Hz) and noisy speech with different SNR's. The following kinds of noise were considered:

- White noise: It has a flat power spectral density (PSD).
- Pink noise: The PSD is inversely proportional to the frequency.
- Blue noise: The PSD is directly proportional to the frequency.
- Babble noise: A mixture of speech from multiple speakers.
- Vehicle noise: Noise from a moving military vehicle.
- HF channel noise: Noise from a high frequency radio channel.

Among the above noises, pink and blue are simulated colored noises, while babble, vehicle and HF channel are real world noises taken from the Noisex database [18]. When only a few subbands are affected by noise, the GCI information in the other bands not affected by noise may be intact. In order to test this, we have created coloured noises that affect low or high frequency subbands differently. Additionally, the effect of all the subbands being affected is studied by using white noise and real world noises.

We have also tested the robustness of our methods to distortions of the speech signal by reverberation, which is the accumulation of reflected sounds from the surfaces in an enclosed space. Reverberant speech $s_r[n]$ is simulated by $s_r[n] = s[n] * r[n]$ where $s[n]$ is the speech signal, $r[n]$ is the room impulse response (RIR). RIR is simulated assuming a room of size $5 \times 4 \times 6$ m for varying values of T60, which is the time required for the magnitude of the RIR to decay by 60 dB. We have used the code for generating RIR available at [161] which uses the image method, proposed by Allen et al. [162].

We have implemented our algorithm as well as simulated the first three noises in MATLAB, while the last three noises are taken from [18]. Large negative extrema in the dEGG signal correspond to the instants of vocal fold closure [154]. Hence, the local minima in the dEGG signal with absolute value exceeding a specified threshold ($1/6^{th}$ of the peak to peak value of dEGG signal) have been taken as the ground truth for GCIs for evaluating our methods. The intrinsic delay between the microphone recording and the EGG signal has been accounted for.

4.3.3 Performance measures

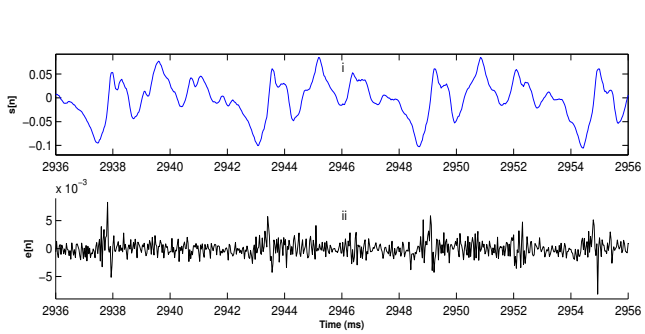
The performance measures are defined with respect to the larynx cycle, which ranges from the mid-point between two adjacent reference GCIs (ground truth) to the next mid-point as defined in [141]. Timing error is defined as the difference between the detected GCI and the reference GCI (from dEGG) in any larynx cycle, where exactly one GCI is detected. The following performance measures have been used for evaluating GCIs:

- Identification rate (IDR): percentage of larynx cycles (PLC) for which exactly one GCI is detected
- Miss rate (MR): PLC for which no GCI is detected
- False alarm rate (FAR): PLC for which more than one GCI is detected
- Standard deviation of error (SDE): standard deviation of the timing error.
- Accuracy to 0.25 ms: PLC for which the absolute value of the timing error is less than 0.25 ms
- Accuracy to 0.50 ms: PLC for which the absolute value of the timing error is less than 0.50 ms

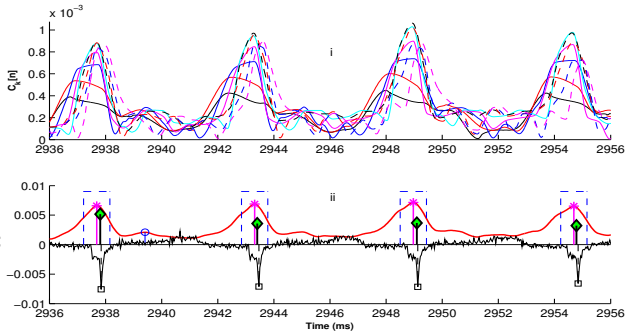
4.4 Simulation results and performance evaluation

4.4.1 Illustration of GCI estimation from clean and noisy speech

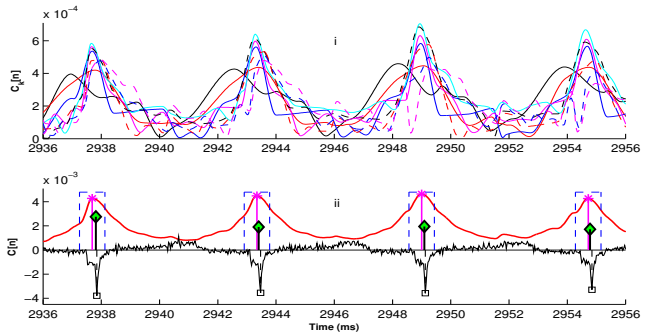
Figure 4.7 shows the estimation of GCIs from the subband components of the LP residual of the clean speech signal extracted using our approach. It can be observed that while the LP residual signal is noisy, peaks of the individual subbands are approximately aligned with the GCIs and hence by adding them, the peaks are retained in the composite signal and closer to the reference GCIs. The final estimated GCIs are obtained as the peaks of the LP residual signal within a rectangular search region around the estimated GCIs.



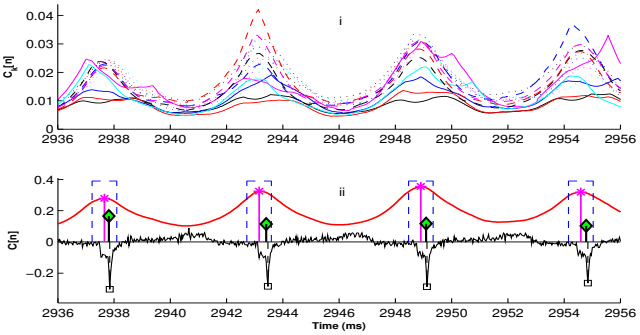
(a) Clean speech signal and its LP residual



(b) Hamming subband envelope (HBE)



(c) Gammatone subband envelope (GTE)



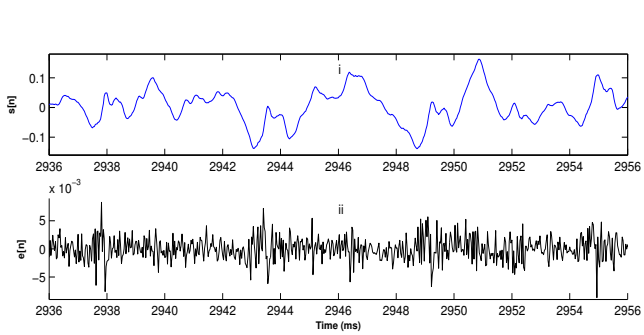
(d) AM-FM decomposition (HAM)

Figure 4.7: Extraction of GCIs from a segment of (clean) voiced speech from SLT database using each of the three proposed methods. (a:i) Speech signal. (a:ii): the LP residual signal. (b,c,d:i) *solid curves* are the first five subbands; *dashed curves* are the next four subbands. (d:i) *dotted curves* are the last five AM components. (b,c,d:ii) GCIs extracted from the composite signal (CS) of subbands. *thick solid curve* is the CS; *circle markers* denote the spurious detections, eliminated later by refinement; *starred markers* denote the initially estimated GCIs, g_{iest}; *dashed rectangle* denote the region around g_{iest}; *diamond markers* denote the final estimated GCIs, g_{est}; *thin solid curve* is the dEGG signal; *square markers* denote the reference GCIs from the dEGG signal.

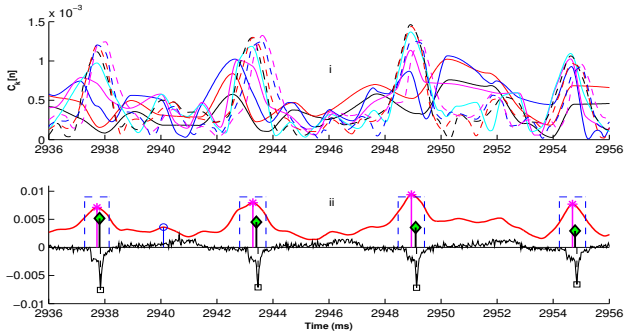
Figure 4.8 shows the intermediate signals involved in the estimation of GCIs for speech with vehicle noise at an SNR of 0 dB, using each of the three proposed methods. It is seen that the GCIs are estimated accurately even in the presence of vehicle noise with mean energy same as the speech signal (0 dB).

4.4.2 Analysis of individual subbands for GCI estimation

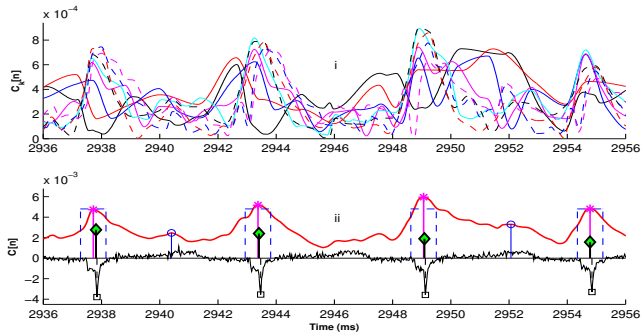
We now study the GCI detection performance using individual subband components. Table 4.1 illustrates how much GCI information is present in individual subbands and the sum of subbands, $C[n]$ by evaluating the performance on the APLAWD database, using measures listed in Sec.4.3.3. GCIs are estimated using individual $C_p[n]$ denoted as HBE 1-9; HBE is the hamming filterbank envelope based method. It is seen from Table 4.1 that IDR for all the individual subbands is above 89% except for HBE 1, which shows that significant GCI information is present in all the subbands. HBE is better in terms of IDR than most individual subbands and has 6% more accuracy to 0.25 ms than the best sub-



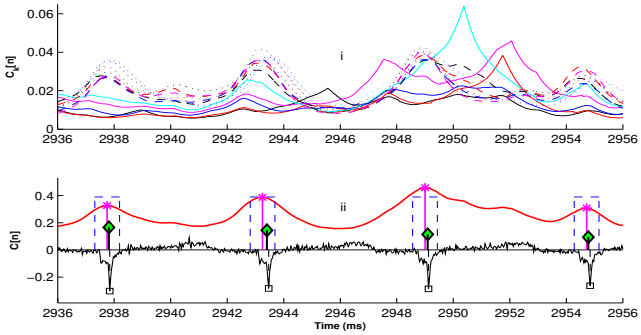
(a) Noisy speech signal and its LP residual



(b) Hamming subband envelope (HBE)



(c) Gammatone subband envelope (GTE)



(d) AM-FM decomposition (HAM)

Figure 4.8: Extraction of GCIs from a segment of voiced speech (SLT database) with additive vehicle noise at SNR of 0 dB using each of the three proposed methods. Notations for the above plots are the same as in Fig. 4.7.

band. The improvement in accuracy to 0.25 ms using HBE over all other individual subbands proves our assumption that GCI information is distributed across different subbands for various frames. Also, peaks in the individual subbands may correspond to GCIs, and composite signal obtained by adding the subbands may preserve the peaks due to the smoothness of the subband envelopes. It is seen that using direct envelope of LP residual, we get a IDR of 80% and accuracy to 0.25 ms of 9.71% only as compared to 93.1% and 81.45% using HBE method on the APLAWD database, which is a significant improvement and justifies the use of subbands.

Further, since speech signal is non-stationary, it may seem that dynamically selecting the best subband component for each time frame may give better estimation of GCIs in terms of detection and accuracy than the summed subband components. So, to see the theoretical best performance a subband based approach can achieve, we have estimated initial GCIs using each of the individual subbands, and then using the reference GCI from the dEGG signal, we picked the estimated GCI from the subband nearest to each GCI reference, which we name as HBEBEST. It has to be noted that refinement of initial GCIs using the LP residual signal is not necessary for HBEBEST, since the best subband with reference to ground truth is chosen. Clearly, HBEBEST denotes the best possible result obtainable using the hamming subband envelope approach, if one can come up with an automated algorithm to

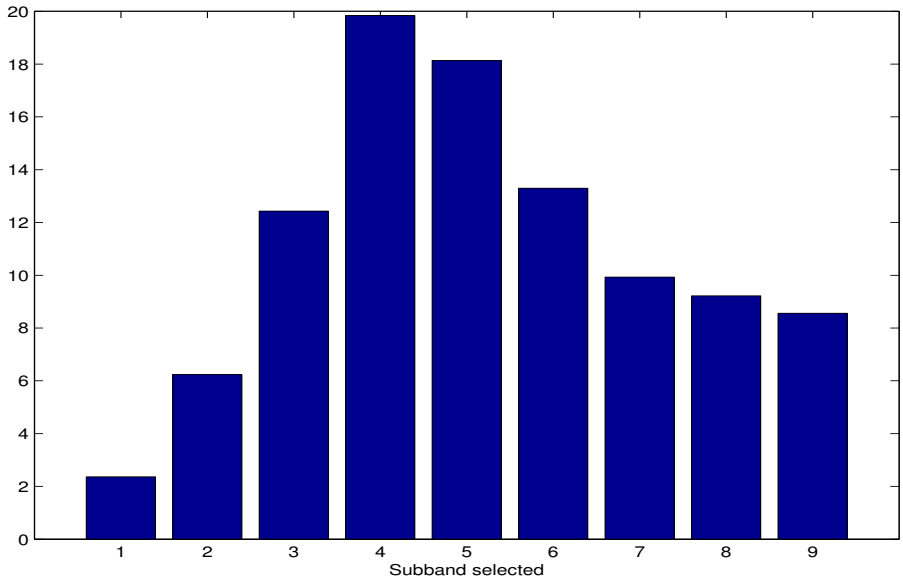


Figure 4.9: Histogram of percentage of GCIs selected as best for each band for clean speech evaluated on the APLAWD database. The spread in the distribution confirms the presence of significant GCI information in each subband.

pick the best subband for each analysis frame. Figure 4.9 shows the histogram of percentage of GCIs from each subband selected as best or closest to the reference GCIs. It is seen from the histogram that all the subbands have significant contribution to the best GCIs and hence, in general, superior GCI estimation performance may be obtained by exploiting the best possible subband everytime.

Table 4.1: Performance comparison of GCI estimation using individual subbands, HBE and HBEBEST on clean speech (from APLAWD database) w.r.t. identification rate (IDR), miss rate (MR), false alarm rate (FAR), accuracy to 0.25 ms and 0.50 ms (Acc.25, Acc.50), all in %. HBEBEST gives the potential performance, if the GCI can be detected from the best possible subband, every time.

Method	IDR	MR	FAR	Acc.25	Acc.50
HBE 1	81.40	18.34	0.26	34.79	61.50
HBE 2	92.69	6.28	1.02	27.19	48.45
HBE 3	94.02	4.93	1.04	29.43	64.71
HBE 4	92.94	5.60	1.46	63.86	72.48
HBE 5	91.95	6.45	1.60	70.71	78.95
HBE 6	92.11	6.38	1.51	74.85	82.40
HBE 7	91.27	7.12	1.61	75.08	81.94
HBE 8	90.36	7.98	1.66	76.13	81.33
HBE 9	89.80	8.47	1.73	75.07	80.59
HBE	93.10	6.29	0.60	81.45	87.26
HBEBEST	98.85	0.56	0.59	89.20	95.45

4.4.3 Performance comparison with other methods

We compare the performance of the three proposed methods with those of ZFF [141], DYPSA [140] and SEDREAMS [142]. We have run the codes available for other methods on the same set of databases and obtained the results.

Figure 4.10 shows the histogram of the GCI timing error for the six methods and HBEBEST on clean speech, over all the databases. Also, SDE and accuracy to 0.25 and 0.5 ms have been shown. In the case of accuracy to 0.25 ms over all the databases, GTE and HBE are 3.7% better than the best of the other methods, while around 2% better for 0.5 ms. It is seen that our methods as well as SEDREAMS are peaky near zero error. HBEBEST is the peakiest near zero error indicating the theoretical potential of subband methods. Table 4.2 compares the six methods on clean and telephone

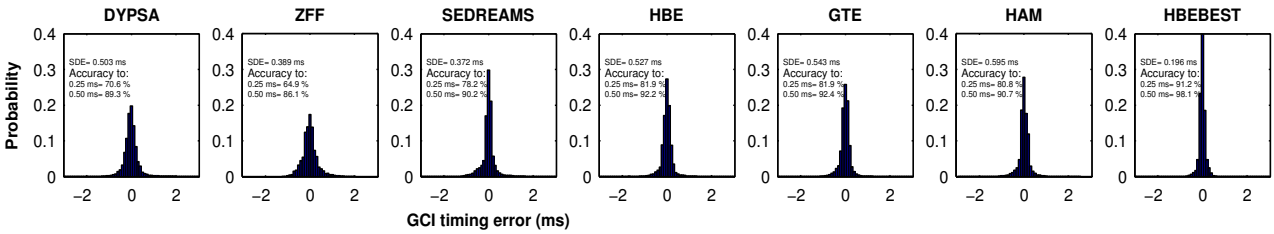


Figure 4.10: Histograms of the GCI timing error of the six methods on clean speech, combining all the databases.

speech for all the databases. For each database, the best performance among the six techniques is highlighted by shading the corresponding cell. Whereas, the bold numbers indicate the cases, where the performance obtained by one of our three methods is the second best.

In general, the performance of our methods is reasonably comparable to the best of the methods in the literature; in some cases, it even surpasses the others in terms of accuracy of detection. Due to the subband analysis in our methods, there is no significant degradation in performance in the case of telephone speech. We have also shown the theoretical best performance which can be achieved using our HBE variant, HBEBEST in a separate italicized row. Ideally, a subband based GCI approach must be able to pick the most accurate GCI from different optimal subband envelopes at different time instants. HBEBEST serves as an upperbound on the performance attainable by any subband approach, since it picks the best GCI estimate at every time instant. Although we note that HBEBEST is of little practical utility, it shows clearly that the subband based approach has potential to estimate GCIs with better accuracy.

It is seen that the identification rates of HBE and HAM method on clean speech are above 97.4% for all the databases except for APLAWD. In the case of ZFF method evaluated on telephone speech, there is a significant degradation in IDR for BDL, RAB and KED databases due to high false detections of GCIs, which also leads to high FAR of above 59%. On telephone speech, our methods give the highest IDR for RAB and KED databases, and lowest FAR.

In the case of accuracy to 0.25 ms, our methods are the best for BDL, JMK and KED databases

Table 4.2: Performance comparison of GCI estimation techniques on clean and telephone speech for all the six databases w.r.t. IDR, MR, FAR, standard deviation of error (SDE in ms), Acc.25 and Acc.50. Shaded cells: best method overall; bold No.: cases, where the performance of one of our methods is the second best.

Database	Method	IDR (%)		MR (%)		FAR (%)		SDE (ms)		Acc. to 0.25 ms (%)		Acc. to 0.50 ms (%)	
		Clean	Telephone	Clean	Telephone	Clean	Telephone	Clean	Telephone	Clean	Telephone	Clean	Telephone
BDL	DYPSA	97.58	96.66	1.28	1.48	1.14	1.86	0.47	0.46	74.77	78.89	91.55	92.37
	ZFF	99.70	40.62	0.21	0.02	0.10	59.36	0.26	0.30	80.93	74.73	95.48	91.89
	SEDREAMS	99.51	93.02	0.33	0.04	0.15	6.94	0.30	0.31	86.30	86.73	92.88	92.25
	HBE	99.02	**98.97**	0.55	0.52	0.43	0.51	0.45	0.47	87.83	88.91	**94.07**	94.09
	GTE	99.12	99.04	0.51	0.53	0.37	**0.43**	0.47	0.47	87.03	**88.42**	93.65	**93.84**
	HAM	98.95	98.92	0.81	0.84	0.24	0.25	0.49	0.50	**87.22**	88.39	93.20	93.38
	HBEBEST	*99.77*	*99.73*	*0.11*	*0.12*	*0.12*	*0.15*	*0.17*	*0.19*	*94.56*	*94.39*	*98.41*	*98.10*
JMK	DYPSA	99.19	97.49	0.41	1.03	0.39	1.48	0.58	0.64	65.46	69.54	86.67	87.53
	ZFF	99.77	98.61	0.22	0.01	0.01	1.38	0.55	0.52	37.27	29.50	67.74	64.77
	SEDREAMS	99.53	99.84	0.41	0.07	0.07	0.09	0.55	0.61	70.43	69.27	84.40	82.25
	HBE	98.09	97.96	0.83	0.89	1.09	1.15	0.71	0.75	77.93	76.91	90.20	**89.60**
	GTE	98.22	98.15	0.78	0.86	0.99	0.99	0.74	0.76	**77.59**	**76.54**	**89.90**	89.66
	HAM	97.46	97.56	1.83	1.76	0.71	**0.67**	0.88	0.86	76.05	69.52	87.63	88.09
	HBEBEST	*99.83*	*99.79*	*0.11*	*0.13*	*0.06*	*0.07*	*0.22*	*0.27*	*90.17*	*86.92*	*98.12*	*96.33*
SLT	DYPSA	98.26	96.81	0.97	1.47	0.77	1.72	0.42	0.41	62.71	62.66	87.85	88.43
	ZFF	99.69	99.40	0.01	0.01	0.29	0.60	0.20	0.19	81.38	82.46	98.02	98.49
	SEDREAMS	99.66	99.48	0.04	0.01	0.31	0.51	0.33	0.32	69.50	67.78	88.63	88.50
	HBE	98.98	98.85	0.39	0.43	0.63	0.72	0.38	0.45	75.51	71.53	92.08	88.95
	GTE	99.01	98.92	0.42	0.42	0.57	0.66	0.37	0.43	**75.76**	**72.87**	**92.63**	**90.56**
	HAM	98.34	98.32	1.31	1.30	0.35	0.38	0.43	0.46	74.46	72.07	90.36	88.99
	HBEBEST	*99.94*	*99.93*	*0.01*	*0.01*	*0.04*	*0.06*	*0.18*	*0.24*	*87.89*	*80.47*	*98.39*	*95.90*
RAB	DYPSA	86.80	87.72	0.47	0.66	12.73	11.62	0.53	0.49	89.85	91.53	95.64	96.14
	ZFF	99.97	22.68	0.03	0.01	0.00	77.31	0.54	0.58	33.47	38.39	64.72	67.68
	SEDREAMS	99.97	96.37	0.02	0.02	0.01	3.61	0.34	0.44	93.55	93.07	95.78	94.97
	HBE	97.54	**97.24**	0.68	0.77	1.78	**2.00**	0.86	0.96	**91.98**	91.40	94.03	93.14
	GTE	96.07	95.12	0.71	0.70	3.22	4.17	0.97	0.97	91.28	91.23	93.26	92.89
	HAM	97.66	97.47	1.01	1.13	1.33	1.40	0.92	0.93	91.02	90.87	93.10	92.69
	HBEBEST	*99.75*	*99.67*	*0.16*	*0.20*	*0.09*	*0.13*	*0.21*	*0.22*	*96.86*	*95.93*	*98.58*	*98.03*
KED	DYPSA	98.90	98.47	0.56	0.71	0.53	0.81	0.38	0.35	89.38	92.78	96.36	96.88
	ZFF	99.75	33.39	0.11	0.07	0.15	66.55	0.62	0.87	34.66	22.28	66.69	53.48
	SEDREAMS	99.92	98.99	0.07	0.00	0.01	1.01	0.36	0.45	89.84	92.90	95.92	94.25
	HBE	99.51	99.44	0.36	0.40	0.13	**0.15**	0.33	0.34	96.20	**96.70**	97.76	**97.79**
	GTE	99.49	**99.39**	0.29	0.34	0.23	0.27	0.33	0.34	96.20	96.78	**97.75**	97.88
	HAM	99.34	99.30	0.58	0.59	**0.08**	0.11	**0.36**	0.36	**95.88**	96.44	97.38	97.50
	HBEBEST	*99.88*	*99.82*	*0.07*	*0.11*	*0.05*	*0.07*	*0.13*	*0.15*	*98.22*	*98.01*	*99.55*	*99.41*
APLAWD	DYPSA	95.21	94.05	3.21	3.39	1.58	2.55	0.69	0.59	73.41	74.12	86.20	87.88
	ZFF	97.19	94.88	2.55	1.45	0.27	3.67	0.47	0.49	57.28	55.16	80.62	77.40
	SEDREAMS	94.82	95.47	5.01	1.58	0.17	2.95	0.35	0.42	83.70	79.96	91.22	88.63
	HBE	93.10	92.81	6.29	6.55	0.60	**0.64**	0.59	0.58	81.45	**79.99**	87.26	86.38
	GTE	93.47	93.51	5.70	5.65	0.83	0.84	0.57	0.58	**83.31**	81.87	**88.81**	87.85
	HAM	87.49	87.25	12.17	12.39	0.34	0.37	0.61	0.62	79.50	78.15	85.63	84.78
	HBEBEST	*98.85*	*98.74*	*0.56*	*0.62*	*0.59*	*0.64*	*0.25*	*0.26*	*89.20*	*88.25*	*95.45*	*94.95*

while they are comparable to the best of the other methods for other databases on both clean and telephone speech. Our methods consistently give above 74% accuracy for all the databases. It is observed that the performance of ZFF varies widely from one database to another, since it gives poor accuracy for JMK, RAB, KED and APLAWD databases compared to other methods.

It is seen that although HBEBEST is not a realistic method, it mostly outperforms all the other methods, in terms of almost all the performance measures. This shows that the subband approach has a promise for the future, to give better and robust GCIs, based on further, in depth studies. All the methods have been implemented using MATLAB on a Intel Core i5 CPU 650, 3.2 GHz with 3 GB of RAM.

Figure 4.11 shows the variation of performance of all the six methods and HBEBEST on speech with six different kinds of additive noises with SNR varying from -5 dB to 30 dB in steps of 5 dB and reverberation with T60 ranging from 150 to 850 ms in steps of 100 ms. The values plotted correspond to the combined evaluation on all the six databases.

It is seen that for white noise, ZFF gives the best performance in terms of IDR, SDE and accuracy, below SNR of 15 dB. At low SNRs, our methods give second best performance for accuracy to 0.25 and 0.5 ms. DYPSA has an IDR poorer than our methods in the case of white, pink and blue noises due its high FAR.

In the case of pink noise, our methods give accuracy comparable to other methods, surpassing other methods at low SNRs,while the IDR of our methods is around 10% higher than SEDREAMS, at very low SNR.

All our methods give better accuracy to 0.25 and 0.5 ms than SEDREAMS and DYPSA in the case of blue noise at all SNRs.

It is observed that ZFF is robust to variation in SNR in the case of blue and HF channel noise as seen from its flat IDR, SDE and accuracy to 0.25 and 0.5 ms, mainly due to its immunity to high frequencies.

In the case of babble noise, our methods give accuracy to 0.25 ms similar to SEDREAMS and outperforms ZFF at all SNRs. It is seen that DYPSA has better accuracy at low SNRs. In the case of vehicle noise, all our methods give better IDR than SEDREAMS and ZFF at low SNR, and they outperform other methods for accuracy to 0.25 ms at all SNRs.

In the case of HF channel noise, our methods outperform DYPSA in the case of IDR.

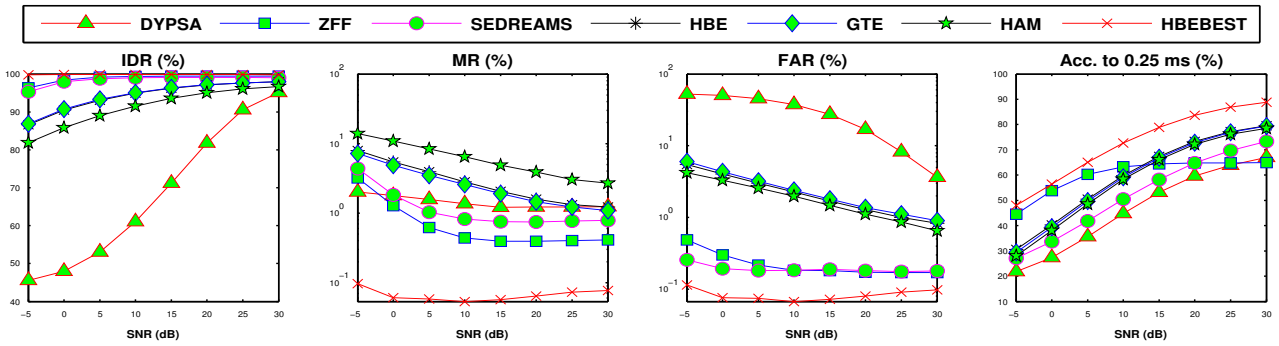
In the case of reverberation, IDR of DYPSA degrades with increasing T60 due to its high FAR. All the methods perform poorly with respect to accuracy to 0.25 and 0.5 ms. Accuracy of DYPSA does not degrade with increasing T60, while our methods perform better than ZFF and SEDREAMS in terms of accuracy.

It is seen that HBE and GTE have similar performaces in all the cases. HBEBEST outperforms all the methods for all performance measures by a considerable margin except in the case of HF channel noise at low SNRs, for which ZFF gives slightly better accuracy. Also, HBEBEST gives much better accuracy to 0.25 and 0.5 ms in the case of reverberation.

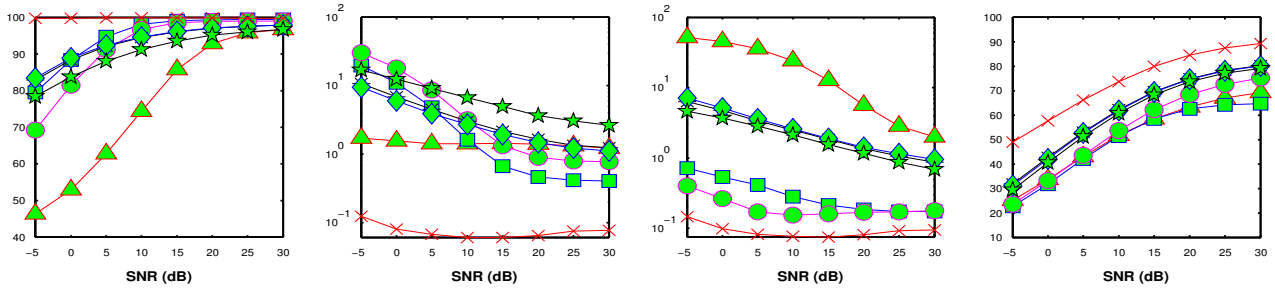
Table 4.3 shows the comparison of HBE with HBEVAR for white, babble and vehicle noises. We have evaluated HBEVAR method on all the databases and found that we get a overall IDR of 98.20% and accuracy to 0.25 ms of 81.5% as compared to HBE which gives an IDR of 98.05% and accuracy to 0.25 ms of 81.9%. It is seen that for white and babble noises, HBEVAR gives around 1-2% better IDR than HBE while accuracy to 0.25 ms remains similar.

4.5 Discussion and Conclusion

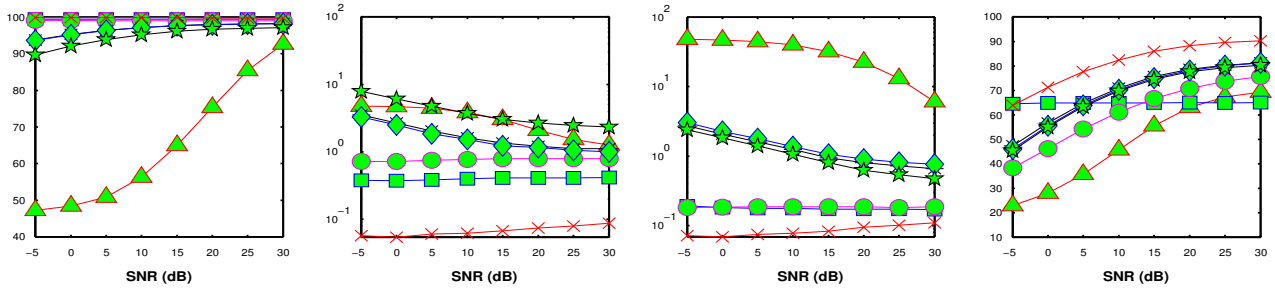
In general, accurate estimation of GCIs in the presence of additive noise is a challenging task. With noisy speech, the summation of subband components retains the manifestations of impulse-like ex-



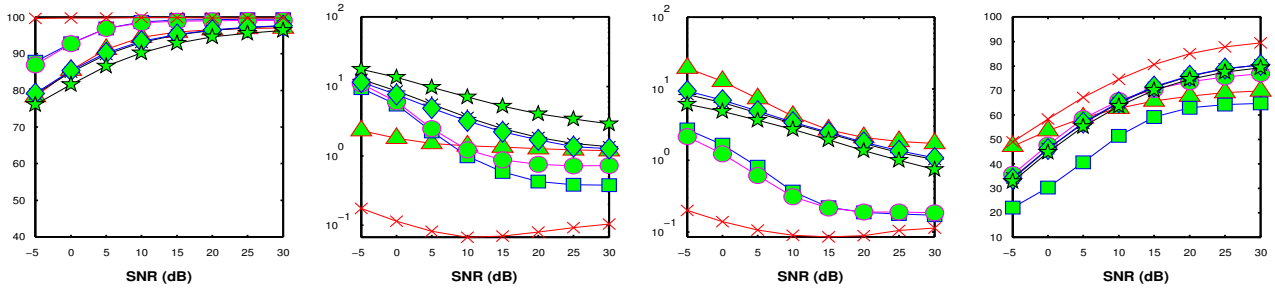
(a) White Noise



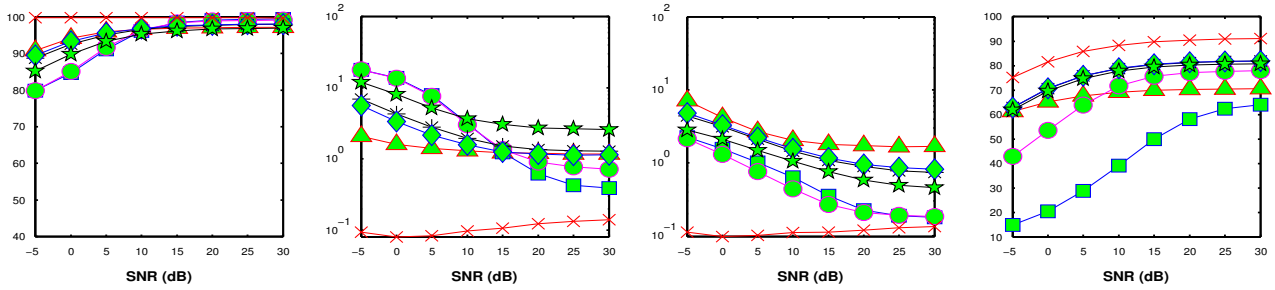
(b) Pink Noise



(c) Blue Noise

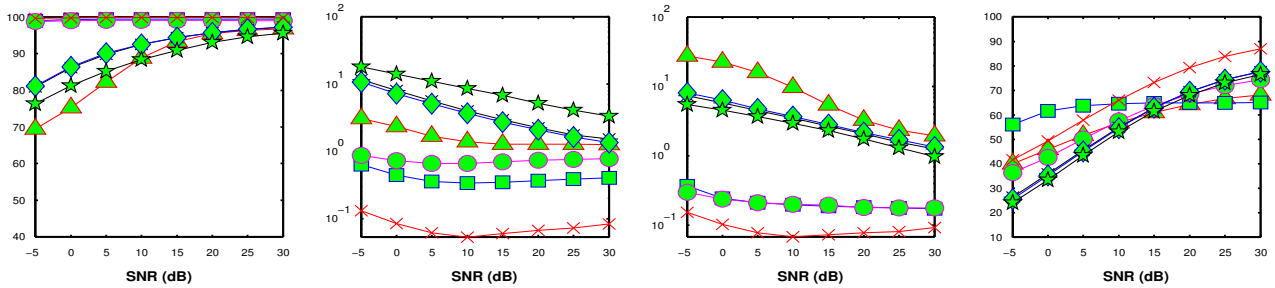


(d) Babble Noise

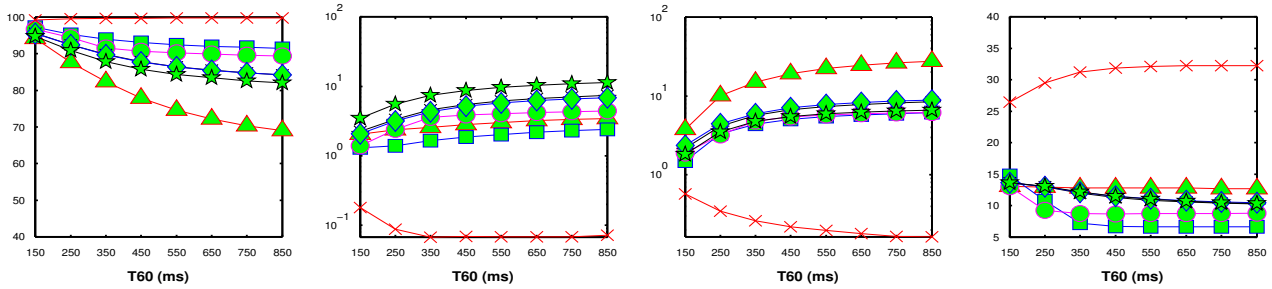


(e) Vehicle Noise

Figure 4.11: (continued)



(a) HF channel Noise



(b) Reverberation

Figure 4.11: Robustness of GCI estimation methods to different kinds of noise and reverberation w.r.t. IDR, MR, FAR and Acc.25 (combined performance on all the databases). Miss rate and false alarm rate are shown in logarithmic scale.

Table 4.3: Evaluation of HBE and HBEVAR methods on the combined databases with additive noises at different SNRs

	Method	Clean	White noise				Babble noise				Vehicle noise			
SNR (dB)			-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
IDR (%)	HBE	98.05	86.64	90.50	93.07	94.94	78.99	84.97	89.87	93.12	88.69	92.57	95.10	96.61
	HBEVAR	98.20	88.83	92.32	94.44	95.95	79.78	86.07	90.99	94.05	88.38	92.65	95.32	96.84
MR (%)	HBE	1.26	7.92	5.46	3.93	2.80	12.68	8.65	5.53	3.59	6.91	4.31	2.80	1.94
	HBEVAR	1.13	7.26	4.93	3.48	2.49	12.86	8.53	5.33	3.35	6.98	4.15	2.62	1.76
FAR (%)	HBE	0.69	5.44	4.04	3.01	2.26	8.34	6.39	4.60	3.29	4.39	3.12	2.10	1.45
	HBEVAR	0.66	3.92	2.75	2.08	1.56	7.37	5.39	3.68	2.60	4.64	3.20	2.06	1.40
SDE (ms)	HBE	0.53	1.28	1.11	0.96	0.84	1.50	1.27	1.07	0.91	1.07	0.89	0.75	0.65
	HBEVAR	0.49	1.17	0.99	0.86	0.75	1.50	1.25	1.02	0.86	1.13	0.91	0.73	0.62
Acc.25 (%)	HBE	81.90	30.55	40.32	50.50	59.81	34.14	46.10	56.49	64.90	62.95	70.63	75.76	78.82
	HBEVAR	81.51	31.57	41.32	51.05	59.95	33.29	45.38	56.01	64.49	60.08	68.91	74.74	78.10

citations of all the components, giving good initial estimates of GCIs. The results show that the performance of the proposed algorithms is comparable to other methods for various kinds of noisy speech, telephone, reverberant and clean speech. We also observe that no single method gives the best performance across all the types of noisy speech at different SNRs.

Initial estimate of GCI derived from the composite signal of the subbands of the LP residual signal is used to narrow down the region for finding the final GCI, since directly estimating GCIs from the LP residual is difficult due to its noisy structure. The good performance of our methods even in the presence of various kinds of additive noise is probably because we exploit the GCI information present

in all the subband components of speech. The composite signal obtains a better estimate of the initial excitation instants even in the presence of noise, and the accuracy degrades gradually with decrease in SNR.

We have clearly shown that significant GCI information exists in each frequency band of speech up to 2000 Hz, and a minimum of 81% IDR can be obtained for clean speech from any subband using the HBE method. Among the proposed methods, HBE and GTE are the most promising for GCI estimation. Also, dynamic selection of the best subband using some additional knowledge may achieve robust GCI estimation closer to the HBEBEST. As an enhancement to this approach, different filterbanks may be explored with varying bandwidths and filter characteristics. Further experiments for selecting the best band may lead to improvement of the performance, in terms of the detection rate and/or accuracy.

Chapter 5

Conclusion and future work

In this chapter, we summarize our findings in this thesis and suggest directions for possible future work.

5.1 Conclusions

Real life speech signals generally contain a foreground speech by a particular speaker in the presence of a background environment like factory or traffic noise. Processing of these signals has been approached by the research community for various independent applications like classification of components of the noisy speech signal, source separation, enhancement, speech recognition, audio coding, duration modification and speaker normalization. Machine listening encapsulates solutions to these applications in a single system. It extracts useful information from noisy speech signals, and attempts to understand the content as much as humans do. We have tried to emulate a machine listening system and have been partially successful to extract different kinds of information from a noisy speech signal. We have dealt with noise and speaker classification in a clean case, noisy speech, overlapped and conversational speech. We have then separated and obtained the enhanced speech. The case when the training data for a particular speaker/noise is not available is tackled by selecting the closest matched dictionaries and using adaptive update. We achieve a reasonable speaker classification accuracy of 75.6% and noise classification accuracy of 99.8% at an SNR of 0 dB. The above classification/separation of noisy speech is useful for hearing aids and mobile communications. Further, we have looked at the detection of transitions in a speech signal at the level of phonetic classes. We have been broadly able to assign specific classes of detected transitions to changes in the speech signal structure from sonorants to non-sonorants and silence to sonorant/non-sonorants and vice versa. In the voiced part of the speech signal, we have looked at the detection of instants at which the glottis closes (GCIs), which is of importance in voice modification and speech synthesis. These GCIs occur quasi periodically giving the information about the variation of the instantaneous pitch frequency in a speech signal.

So, starting at the broad classification of components in an audio signal like speaker and noise sources, we have looked at the intricate structure of the speech signal and segmented it into different

regions based on changes in the signal and source characteristics. We have achieved good results on various aspects of speech analysis like noise and speaker classification in noisy/overlapped speech, source separation performance, perceptual evaluation of source separation, detection of transitions and GCIs. There is a scope for improvement and to make the system more practically feasible as discussed in the next section on scope for future work.

5.2 Scope for future work

In this thesis, the sparse representation based approach for classification and separation of the noisy speech has been dealt separately from the acoustic-phonetics knowledge based approach for detection of transitions and GCIs. As a future work, we can combine both the approaches and build a unified framework for classification, separation and detection of transitions between class of phones. Also, some of the information extracted using acoustic-phonetics knowledge based approach for the detection of significant transitions robust to various noises may be useful for better classification and separation of noisy speech signals. For example, pitch synchronous frames extracted by using the segment between two successive GCIs can be used for learning dictionaries which have better structure than fixed frame-size dictionaries and may give better source separation. Also, learning phonetic class specific dictionaries may be useful for speech recognition and detection of co-occurrence of phones in an overlapped speech. We can learn block mag.STFT features (feature spanning multiple frames for a given frequency range) so that it captures temporal variation across both frequency and time in the same feature vector.

Also, in this thesis, we have addressed noisy speech where maximum of two sources are active at the same time. We can generalize the problem to K unknown No. of sources active at the same time where K_{sp} is the unknown No. of speaker sources and K_{ns} is the unknown No. of noise sources such that $K = K_{sp} + K_{ns}$. In this case, we have to estimate K_{sp} and K_{ns} No. of speaker and noise sources present, which may vary from one noisy segment to another across time. Then, given the dictionaries for multiple speaker and noise sources, we can separate the noisy speech and obtain K streams of separated speech/noise components.

The sparse representation based classification can be extended to the identification of language [163], gender, accent, dialect, emotions and other classes commonly incurred in an audio signal. The algorithms proposed in this thesis for classification and separation of noisy speech may be converted to a module useful in hearing aids and mobile communications.

Bibliography

[1] C. Müller, *Speaker Classification II*. Springer, 2007. 1

[2] C. Couvreur, V. Fontaine, P. Gaunard, and C. G. Mubikangiey, "Automatic classification of environmental noise events by hidden Markov models," *Applied Acoustics*, vol. 54, no. 3, pp. 187–206, 1998. 1, 11

[3] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007. 1, 11, 20

[4] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005. 1, 45

[5] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech communication*, vol. 49, no. 7, pp. 588–601, 2007. 1, 10

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013. 1, 45

[7] L. R. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993. 1

[8] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010. 1, 11

[9] J. Nikunen and T. Virtanen, "Object-based audio coding using non-negative matrix factorization for the spectrogram representation," in *Audio Engineering Society Convention 128*, Audio Engineering Society, 2010. 1, 11

[10] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, 2006. 1

[11] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Acoustics, Speech, and Signal Processing (ICASSP), 1996 IEEE International Conference on*, vol. 1, pp. 353–356, IEEE, 1996. 1

[12] R. G. Malkin, *Machine listening for context-aware computing.* PhD thesis, Carnegie Mellon University Pittsburgh, PA, 2006. 1, 11

[13] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in *Proc. of the 6th Int. Conference on Digital Audio Effects*, Citeseer, 2003. 1

[14] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis.* New York, NY, USA: Elsevier Science Inc., 1995. 2

[15] T. Shameem, "The organs of speech." http://tanvirdhaka.blogspot.in/2010/12/organs-of-speech.html. [Online] Accessed: 2017-03-30. xviii, 2

[16] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990. 2

[17] "Voiced and unvoiced speech overview." http://www.seas.ucla.edu/dsplab/vus/over.html. [Online] Accessed: 2017-03-30. 2

[18] "NOISEX-92." http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html. [Online] Accessed: 2017-03-30. xxiii, 4, 17, 86, 100, 101

[19] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer Publishing Company, Incorporated, 1st ed., 2010. 5

[20] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004. 5, 14

[21] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993. 5

[22] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pp. 40–44, IEEE, 1993. 5

[23] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001. 5

[24] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997. 5

[25] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active-set Newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, 2013. 5, 10, 11, 12, 15, 16

[26] T. Virtanen, B. Raj, J. F. Gemmeke, *et al.*, "Active-set Newton algorithm for non-negative sparse coding of audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3092–3096, IEEE, 2014. 5, 15, 16, 24

[27] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, p. 607, 1996. 5

[28] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000. 5

[29] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural networks: Tricks of the trade*, pp. 561–580, Springer, 2012. 5

[30] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003. 5

[31] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-frame compression: Theory and design," *Signal Processing*, vol. 80, no. 10, pp. 2121–2140, 2000. 5

[32] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006. 5, 13

[33] W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (SimCO) for dictionary update and learning," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6340–6353, 2012. 5

[34] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1025–1031, 2011. 5

[35] S. Kong and D. Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," *Computer Vision–ECCV 2012*, pp. 186–199, 2012. 5

[36] S. Shafiee, F. Kamangar, V. Athitsos, and J. Huang, "The role of dictionary learning on sparse representation-based classification," in *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, p. 47, ACM, 2013. 5

[37] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Independent Component Analysis and Signal Separation*, pp. 414–421, 2007. 5

[38] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 989–1001, 2011. 6

[39] D. El Badawy, A. Ozerov, and N. Q. Duong, "Relative group sparsity for non-negative matrix factorization with application to on-the-fly audio source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 256–260, IEEE, 2015. 6

[40] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal processing*, vol. 81, no. 11, pp. 2353–2362, 2001. 6

[41] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010. 6, 11, 20

[42] K. N. Stevens, *Acoustic phonetics*, vol. 30. MIT press, 2000. 6

[43] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report*, vol. 93, 1993. xxiii, 6, 7, 17, 18, 72, 81

[44] K. V. V. Girish, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources," in *India Conference (INDICON), 2016 IEEE Annual*, IEEE, 2016. 8, 13

[45] K. V. V. Girish, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation," in *INTERSPEECH*, 2016. 8, 12, 39

[46] K. V. V. Girish, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Adaptive dictionary based approach for background noise and speaker classification and subsequent source separation," *arXiv preprint arXiv:1609.09764*, 2016. 8

[47] T. V. Ananthapadmanabha, K. V. V. Girish, and A. G. Ramakrishnan, "Detection of transitions between broad phonetic classes in a speech signal," *arXiv preprint arXiv:1411.0370*, 2014. 8

[48] V. R. Lakkavalli, K. V. V. Girish, S. Harshavardhan, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Subband analysis of linear prediction residual for the estimation of glottal closure instants," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 945–949, IEEE, 2014. 8

[49] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 2834–2838, IEEE, 2013. 9

[50] D. Baby, T. Virtanen, T. Barker, *et al.*, "Coupled dictionary training for exemplar-based speech enhancement," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 2883–2887, IEEE, 2014. 10

[51] R. Turner, "Noises off: the machine that rubs out noise." http://www.eng.cam.ac.uk/news/noises-machine-rubs-out-noise-0. [Online] Accessed: 2017-04-12. 10

[52] S. Ikram and H. Malik, "Digital audio forensics using background noise," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pp. 106–110, IEEE, 2010. 10

[53] R. H. Lyon, *Machinery noise and diagnostics*. Butterworth-Heinemann, 2013. 10

[54] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 885–888, IEEE, 2006. 10

[55] A. Shirkhodaie and A. Alkilani, "A survey on acoustic signature recognition and classification techniques for persistent surveillance systems," in *SPIE Defense, Security, and Sensing*, pp. 83920U–83920U, International Society for Optics and Photonics, 2012. 10

[56] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and

efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010. 10, 36

[57] J. M. Kates, "Classification of background noises for hearing-aid applications," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, 1995. 11

[58] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 237–240, IEEE, 1999. 11, 33

[59] M. A. Casey, "Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," in *Workshop for Consistent & Reliable Acoustic Cues*, p. 167, 2001. 11

[60] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009. 11, 33

[61] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden markov model," in *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 27–32, IEEE, 1998. 11

[62] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on speech and audio processing*, vol. 9, no. 4, pp. 441–457, 2001. 11

[63] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002. 11

[64] L. Ma, D. Smith, and B. P. Milner, "Context awareness using environmental noise classification," in *Eurospeech*, 2003. 11

[65] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 3, no. 2, pp. 1–22, 2006. 11

[66] S. Cherla and V. Ramasubramanian, "Audio analytics by template modeling and 1-pass DP based decoding," in *INTERSPEECH*, 2010. 11

[67] V. Ramasubramanian, R. Karthik, S. Thiyagarajan, and S. Cherla, "Continuous audio analytics by HMM and viterbi decoding," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 2396–2399, IEEE, 2011. 11

[68] V. Ramasubramanian, S. Thiyagarajan, G. Pradnya, H. Claussen, and J. Rosca, "Two-class verifier framework for audio indexing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 838–842, IEEE, 2013. 11

[69] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pp. 1–4, IEEE, 2013. 11, 33

[70] B. Cauchi, "Non-negative matrix factorisation applied to auditory scenes classification," *Master's thesis, Master ATIAM, Université Pierre et Marie Curie*, 2011. 11

[71] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pp. 1–9, IEEE, 2013. 11

[72] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1827–1837, 2013. 11

[73] C. Tzagkarakis and A. Mouchtaris, "Sparsity based robust speaker identification using a discriminative dictionary learning approach," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pp. 1–5, IEEE, 2013. 11, 12

[74] C. Joder and B. Schuller, "Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition," in *Speech Communication; 10. ITG Symposium; Proceedings of*, pp. 1–4, VDE, 2012. 11, 44

[75] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 140–148, Springer, 2010. 11, 20

[76] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011. 11

[77] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *INTERSPEECH*, pp. 717–720, 2010. 11

[78] Y.-C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327–1336, 2005. 11

[79] S. Zubair, F. Yan, and W. Wang, "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Digital Signal Processing*, vol. 23, no. 3, pp. 960–970, 2013. 11

[80] J. Le Roux, F. Weninger, and J. R. Hershey, "Sparse NMF–half-baked or well done?," *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015. 11, 13, 15, 18, 24, 30

[81] A. Lefevre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 21–24, IEEE, 2011. 13

[82] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 4–7, 2010. 13

[83] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Optics & Photonics 2005*, pp. 591411–591411, International Society for Optics and Photonics, 2005. 13

[84] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel, "K-WEB: Nonnegative dictionary learning for sparse image representations," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pp. 146–150, IEEE, 2013. 13

[85] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 657–664, 2013. 13

[86] I. Ramírez, F. Lecumberry, and G. Sapiro, "Universal priors for sparse modeling," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2009 3rd IEEE International Workshop on*, pp. 197–200, IEEE, 2009. 13

[87] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2055–2065, 2013. 13

[88] D. Barchiesi and M. D. Plumbley, "Learning incoherent subspaces: classification via incoherent dictionary learning," *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 189–199, 2015. 13

[89] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013. 13

[90] H. Zheng and D. Tao, "Discriminative dictionary learning via Fisher discrimination K-SVD algorithm," *Neurocomputing*, vol. 162, pp. 9–15, 2015. 13

[91] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, "Algorithms, initializations, and convergence for the nonnegative matrix factorization," *arXiv preprint arXiv:1407.7299*, 2014. 16

[92] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009. 18

[93] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006. 23, 26

[94] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994. 44, 45

[95] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Noise robust speaker recognition with convolutive sparse coding," in *INTERSPEECH*, pp. 244–248, 2015. 45

[96] "Veena music and traffic noise." `http://www.youtube.com`. [Online] Accessed: 2014-03-30. 54

[97] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on selected areas in communications*, vol. 10, no. 5, pp. 819–829, 1992. 55

[98] G. Fant, "Speech sounds and features," 1973. 65

[99] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, *et al.*, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Acoustics, Speech, and Signal Processing (ICASSP), 2005 IEEE International Conference on*, vol. 1, pp. I–213, IEEE, 2005. 65, 85

[100] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014. 65, 66

[101] A. K. V. SaiJayram, V. Ramasubramanian, and T. V. Sreenivas, "Robust parameters for automatic segmentation of speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–513, IEEE, 2002. 65, 66

[102] J. P. van Hemert, "Automatic segmentation of speech," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 1008–1012, 1991. 65, 66

[103] R. Andre-Obrecht, "Automatic segmentation of continuous speech signals," in *Acoustics, Speech, and Signal Processing (ICASSP), 1986 IEEE International Conference on*, vol. 11, pp. 2275–2278, IEEE, 1986. 65, 66

[104] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *Acoustics, Speech, and Signal Processing (ICASSP), 1987 IEEE International Conference on*, vol. 12, pp. 77–80, IEEE, 1987. 65, 66

[105] A. Sarkar and T. V. Sreenivas, "Automatic speech segmentation using average level crossing rate information," in *Acoustics, Speech, and Signal Processing (ICASSP), 2005 IEEE International Conference on*, vol. 1, pp. I–397, IEEE, 2005. 65

[106] G. Ananthakrishnan, H. Ranjani, and A. G. Ramakrishnan, "Language independent automated segmentation of speech using Bach scale filter-banks," in *Intelligent Sensing and Information Processing, 2006. ICISIP 2006. Fourth International Conference on*, pp. 115–120, IEEE, 2006. 65

[107] F. Jelinek, L. Bahl, and R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transactions on Information Theory*, vol. 21, no. 3, pp. 250–256, 1975. 66

[108] D. T. Toledano, L. A. H. Gómez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE transactions on speech and audio processing*, vol. 11, no. 6, pp. 617–625, 2003. 66

[109] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997. 66

[110] W. J. Barry and W. A. Van Dommelen, *The integration of phonetic knowledge in speech technology*, vol. 25. Springer Science & Business Media, 2006. 66

[111] P. Niyogi and P. Ramesh, "The voicing feature for stop consonants: Recognition experiments with continuously spoken alphabets," *Speech Communication*, vol. 41, no. 2, pp. 349–367, 2003. 66, 88

[112] A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on

phonetic features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1154–1168, 2008. 66, 67

[113] R. Jakobson, C. G. Fant, and M. Halle, "Preliminaries to speech analysis. the distinctive features and their correlates.," 1951. 66

[114] N. Chomsky and M. Halle, "The sound pattern of english.," 1968. 66

[115] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000. 66, 67

[116] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features.," in *INTERSPEECH*, 2002. 66

[117] O. Scharenborg, V. Wan, and R. K. Moore, "Capturing fine-phonetic variation in speech through automatic classification of articulatory features," in *Speech Recognition and Intrinsic Variation Workshop*, 2006. 66

[118] I. Bromberg, Q. Qian, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, *et al.*, "Detection-based ASR in the automatic speech attribute transcription project.," in *INTERSPEECH*, pp. 1829–1832, 2007. 66

[119] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," *Computer Speech & Language*, vol. 21, no. 4, pp. 620–640, 2007. 67

[120] A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in *Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on*, vol. 2, pp. 726–730, IEEE, 2002. 67, 85

[121] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002. 67, 68, 83

[122] T. Pruthi and C. Espy-Wilson, "Automatic classification of nasals and semivowels," in *International Conference on Phonetic Sciences, Barcelona, Spain*, Citeseer, 2003. 67

[123] S. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information.," in *INTERSPEECH*, pp. 1133–1136, 2005. 67, 85

[124] S. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009. 67, 85, 86, 88

[125] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 299–302, 2013. 67

[126] C. Espy-Wilson, "A phonetically based semivowel recognition system," in *Acoustics, Speech, and Signal Processing (ICASSP), 1986 IEEE International Conference on*, vol. 11, pp. 2775–2778, IEEE, 1986. 67

[127] Z. Zhang, C. Y. Espy-Wilson, and M. Tiede, "Acoustic modeling of American English lateral approximants.," in *INTERSPEECH*, 2003. 67

[128] A. M. Abdelatty Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2217–2235, 2001. 67

[129] N. N. Bitar, *Acoustic analysis and modeling of speech based on phonetic features*. Boston University, 1998. 67

[130] T. V. Ananthapadmanabha, A. P. Prathosh, and A. G. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 460–471, 2014. 67, 83

[131] N. Dhananjaya, B. Yegnanarayana, and P. Bhaskararao, "Acoustic analysis of trill sounds," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3141–3152, 2012. 67

[132] A. Salomon, C. Y. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1296–1305, 2004. 67, 83, 85, 88

[133] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, 1996. 67, 70, 83, 85, 88

[134] D. R. Reddy, "Phoneme grouping for speech recognition," *The Journal of the Acoustical Society of America*, vol. 41, no. 5, pp. 1295–1300, 1967. 67

[135] P. Niyogi and M. Sondhi, "Detecting stop consonants in continuous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1063–1076, 2002. 86

[136] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992. 88

[137] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, no. 2, pp. 143–154, 2004. 91

[138] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 309–319, 1979. 91

[139] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, 1995. 91

[140] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–349, IEEE, 2002. 91, 105

[141] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008. 91, 101, 105

[142] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals.," in *INTERSPEECH*, pp. 2891–2894, 2009. 91, 105

[143] N. Sturmel, C. d'Alessandro, and F. Rigaud, "Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform," in *Acoustics, Speech, and Signal Processing (ICASSP), 2009 IEEE International Conference on*, pp. 4517–4520, IEEE, 2009. 91

[144] S. C. Sekhar, S. Pilli, C. Lakshmikanth, and T. V. Sreenivas, "Novel auditory motivated subband temporal envelope based fundamental frequency estimation algorithm," in *Signal Processing Conference, 2006 14th European*, pp. 1–5, IEEE, 2006. 91

[145] K. Gopalan, "Pitch estimation using a modulation model of speech," in *Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on*, vol. 2, pp. 786–791, IEEE, 2000. 91

[146] C. Prakash, N. Dhananjaya, and S. V. Gangashetty, "Detection of glottal closure instants from Bessel features using am-fm signal," in *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*, pp. 1–4, IEEE, 2011. 91

[147] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012. 91

[148] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013. 91

[149] J. P. Cabral, J. Kane, C. Gobl, and J. Carson-Berndsen, "Evaluation of glottal epoch detection algorithms on different voice types," in *INTERSPEECH*, pp. 1989–1992, 2011. 91

[150] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012. 91, 99

[151] V. R. Lakkavalli, K. V. V. Girish, and A. G. Ramakrishnan, "Sub-band envelope approach to obtain instants of significant excitation in speech," in *Communications (NCC), 2012 National Conference on*, IEEE, 2012. 92

[152] A. Bouzid and N. Ellouze, "Open quotient measurements based on multiscale product of speech signal wavelet transform," *Journal of Electrical and Computer Engineering*, vol. 2007, 2008. 92

[153] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE transactions on Information Theory*, vol. 38, no. 2, pp. 917–924, 1992. 92

[154] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995. 97, 101

[155] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980. 97

[156] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech communication*, vol. 28, no. 3, pp. 195–209, 1999. 98

[157] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993. 98

[158] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Acoustics, Speech, and Signal Processing (ICASSP), 1990 International Conference on*, pp. 381–384, IEEE, 1990. 98

[159] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988. 99

[160] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004. 99

[161] "Room impulse response generator for MATLAB." http://home.tiscali.nl/ehabets/rir_generator.html. [Online] Accessed: 2014-03-30. 101

[162] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979. 101

[163] K. V. V. Girish, V. Vijai, and A. G. Ramakrishnan, "Relationship between spoken indian languages by clustering of long distance bigram features of speech," in *India Conference (INDICON), 2016 IEEE Annual*, IEEE, 2016. 112