ROBUST RECOGNITION OF BINAURAL SPEECH SIGNALS USING TECHNIQUES BASED ON HUMAN AUDITORY PROCESSING

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Anjali I. Menon

B.Tech, Electronics and Communication, Nirma University
M.S., Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

Carnegie Mellon University
Pittsburgh, PA

May 2019

For my heros, Amma and Achan

Acknowledgements

Firstly, I would like to thank my advisor, Prof. Richard Stern. His guidance and support have kept me going through these years. Apart from being a great mentor, he is an incredibly kind person and I am proud to have been one of his students. I'd also like to thank my thesis committee members, Prof. Bhiksha Raj, Dr. Amir Moghimi and Prof. Christopher Brown for their guidance and help. I am particularly grateful to Prof. Bhiksha Raj for his inspirational words that have helped me through difficult times. Special thanks to Dr. Amir Moghimi for supporting me in so many ways as a fellow student first, then as my manager during an internship and now as a thesis committee member. I'd also like to thank Prof. Rita Singh for her guidance throughout this journey. Thanks also to Dr. Chanwoo Kim for his immense help with this thesis.

I am extremely grateful to have gotten a chance at Carnegie Mellon to work and collaborate with some of the most accomplished researchers in their fields and I'd like to especially thank Prof. Alan Black and Prof. Pulkit Grover for giving me a chance to work on some very interesting projects. My sincerest thanks also to the administrative staff at the ECE Department for doing everything in their power to help me out more than once.

There have been many who have honed my interest in this specific field of research and I'd like to thank Prof. Jont Allen and Dr. Mead Killion in particular. My time at the University of Illinois, Urbana-Champaign and at Etymotic research have been a big reason behind my academic pursuits.

I am also very grateful for the friends I have made during graduate school. Thanks to the Hobart Street gang - Meghana, Sunayana, Bhavana and Ruta for all the laughs, the bad movies and for being so generous and helpful. Thanks also to Prasanna, Pallavi, Minhee and Arun for being such amazing friends and for your constant encouragement. There are too many other friends and family members to mention here whose role in my life is

much appreciated but I'd especially like to thank Divya and Kahani for their enduring friendship.

I am blessed to belong to a family that has supported me in all my pursuits. I'd like to thank my parents for teaching me the importance of education and for giving me confidence in my abilities. I have had access to amazing opportunities because of conscious efforts on their part. They have put everything on hold to help me out including during my thesis defense and I know that there is nothing I would have achieved without their love and support. I'd also like to thank my brother Ajay, for being my favourite source of comfort and entertainment. Our conversations about books, tv shows and family gossip kept me informed of the world outside of graduate school . Thanks also to my sister Pratibha for looking out for me all this time.

I am also very lucky to have married into a family that supports and encourages me. I'd like to thank my in-laws for their encouragement all these years and especially for helping me out with finishing up the PhD. A quick shout-out to my brother-in-law Anshit for giving me an opportunity to de-stress-by-celebration, by marrying the lovely Megha.

I'd like to thank my husband, Ankit, without whom none of this would have been possible. He has had to make many sacrifices for my sake and he has done so without hesitation. The PhD process for me, as for most people, has been a test of tenacity and I am glad I had someone with me to dust off the fluff and push me to move on in the face of numerous disappointments. I know I have been unusually lucky to find a board game buddy, a co-vocalist, a fierce cheerleader and my best friend in the same person.

And finally, I'd like to thank Kabir and Ivaan. Their giggles are the sole reason I survive dangerous levels of sleep deprivation. Thank you for all the joy and excitement you have brought into my life.

Thesis Committee

Prof. Richard Stern (**Committee Chair**),

Department of Electrical and Computer Engineering,

Carnegie Mellon University.

Prof. Bhiksha Raj,

Language Technologies Institute,

Carnegie Mellon University.

Prof. Christopher Brown,

School of Health and Rehabilitation Sciences,

University of Pittsburgh.

Dr. Amir Moghimi,

Polycom Inc.

Acknowledgment of sources of financial support

- IARPA BABEL program
- Honeywell Corporation
- Prabhu and Poonam Goel Graduate Fellowship Fund
- Jack and Mildred Bowers Scholarship in Engineering

Abstract

Automatic Speech Recognition (ASR) engines are extremely susceptible to noise. There is an increasing prevalence of voice-assisted devices which need to recognize speech accurately in a variety of complex listening environments. These include the presence of background noise, reverberation, and multiple talkers.

The human auditory system, on the other hand, is very good at understanding speech even in extremely challenging environments. It might therefore, be useful to use our knowledge of human hearing to develop techniques that lead to robust speech recognition. This entails applying techniques that have their basis in human auditory processing towards automatic speech recognition (ASR).

In this thesis, we discuss a number of techniques that address the problem of robust recognition of binaural signals in the presence of reverberation and multiple talkers since they pose a significant problem in terms of ASR engine performance. The techniques discussed here roughly follow the manner in which the auditory system achieves noise robustness. The fundamental idea behind all the techniques proposed is that sounds emanating from the same sound source exhibit some degree of coherence. We aim to use this property to achieve better isolation of the target signal leading to better speech recognition accuracy.

Three techniques are proposed. The Interaural Cross-correlation-based Weighting (ICW) algorithm looks for coherence across sensors using signal envelopes in order to isolate signals coming from the same location. To reduce the effect of reverberation, steady-state suppression is applied as an initial step. The ICW algorithm combined with steady-state suppression leads to significant improvements in ASR accuracy. The Coherence-to-Diffuse Ratio-based Weighting (CDRW) algorithm uses a model-based technique to evaluate the ratio of coherent energy to diffuse energy in a given signal. This leads to significantly better

performance in ASR. The third technique is the Cross-Correlation across Frequency (CCF) algorithm, which looks for coherence in frequency for signal separation. The CCF algorithm also effectively smooths the signal. This algorithm has been tested in conjunction with steady-state suppression and ITD-based analysis. The CCF algorithm leads to improvements in ASR especially in the presence of moderate to high reverberation when the system is trained on clean speech. All algorithms were tested using DNN-based acoustical models obtained with the Kaldi speech recognition toolkit, using both clean and multi-style training data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In recent times the use of speech-driven devices and applications is on the rise. This has been especially true due to the widespread use of home or personal assistants that have a voice-based interface. Such devices have certainly made life much easier. However, the problem of robust speech recognition has become even more pertinent for this reason. Machines or devices using a voice interface are required to work seamlessly in a variety of challenging environments. Some of the important sources of degradation include the presence of noise from the television, cars, street as well as room reverberation.

Speech recognition systems have undergone significant improvements in recent times especially with the advent and widespread use of machine learning techniques [1, 2]. Nevertheless, noise robustness remains problematical, especially if the training data differs significantly from the test cases. Improving speech recognition accuracy in the presence of non-stationary noise sources and other adverse conditions such as reverberation remains a challenge.

The human auditory system, on the other hand, is extremely robust. Listeners can correctly understand speech even in very difficult acoustic environments. This includes the presence of multiple speakers, background noise and reverberation.

It is useful to understand the reason behind the robustness of human perception and

to apply auditory processing-based techniques to improve recognition in noisy and reverberant environments. There have been several successful techniques born out of this approach ($e.g.$ [3, 4, 5, 6, 7] among other sources).

The overall purpose of this thesis is to develop a set of algorithms based on our understanding of the human auditory system that help improve the recognition of speech by Automatic Speech Recognition (ASR) engines in complex acoustical environments. There are a multitude of challenging acoustic environments that a speech recognition system might encounter. In this work, we choose to address the impact of reverberation and interfering talkers in particular.

The rest of this work is organized as follows. Chapter 2 reviews some background information and studies. Chapters 3 provides a detailed explanation of the problem setup and the various systems used for data generation and speech recognition. Chapters 4, 5 and 6 introduce new techniques to achieve better recognition in complex environments. These technique are binaural in nature (for the most part) and are loosely based on human auditory processing. In Chapter 7 we present a comparison of the various techniques detailed in this thesis. Chapter 8 summarizes the primary findings of this thesis and potential future work.

CHAPTER 2

BACKGROUND

2.1 Some basic binaural phenomena

The ability to localize sounds is extremely important and requires binaural cues. The most important cues used by the human auditory system for source localization are the Interaural Time Difference (ITD) and Interaural Intensity Difference (IID). An interaural time difference is produced because it takes longer for a sound to arrive at the ear that is farther away from the source, as seen in Figure 2.1. Similarly, a difference in sound level between the two ears occurs because (at least at higher frequencies) the head partially blocks the propagation of sounds from the source to the ear that is farther from the sound source. This can be measured as an interaural intensity difference.

The effectiveness of ITD and IID cues depend on frequency. The computation of ITD-based cues can be thought of as a comparison of phase between the signals arriving at both the ears. At low frequencies, especially below 1500 Hz, this comparison of phase leads to accurate timing information. However, at higher frequencies, path-length difference of the signals propagated to the two ears can become sufficiently long that the phase delay exceeds half a period. At this point the ITD can no longer be estimated unambiguously.

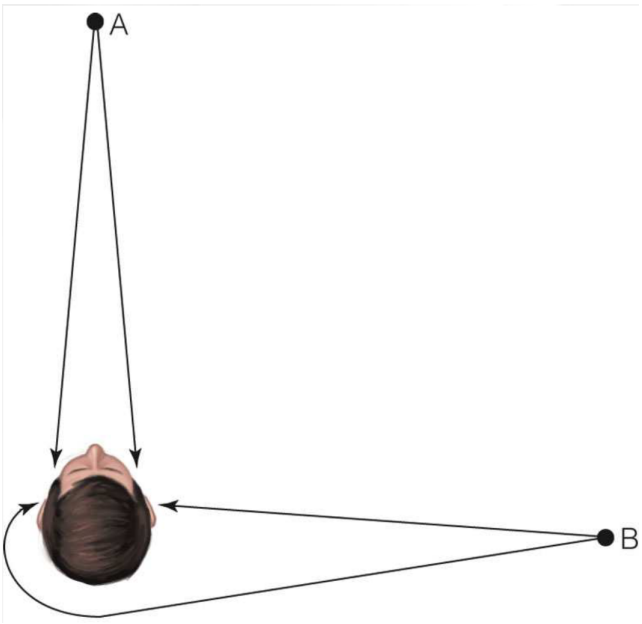


Figure 2.1: Figure illustrating the difference in arrival time between the two ears. Since the sound source A is directly in front of the listener, the sound emanating from A reaches both ears at the same time leading to an ITD of zero. On the other hand, in the case of sound source B, there is a difference in arrival times as well as intensity at both ears. The sound emanating from B reaches the listener's right ear first leading to a non-zero ITD. The intensity of the sound at the right ear will also be slightly higher than the left ear.

This issue is mitigated to some extent by the fact that temporal fine structure information is lost above around 1500 Hz (for humans) and the only timing information that remains available is related to the low-frequency amplitude envelope of the signal. Because of this, the accuracy of the timing information extracted reduces significantly. For this reason, amplitude envelopes are important for ITD-based computations at higher frequencies.

In the case of IIDs, high-frequency content is more important. Low-frequency sounds have a wavelength that is long compared to the size of the head and so the sound is able to diffract around the head. The wavelengths of higher-frequency components, on the other hand, are smaller than the size of the head, which causes these components to be reflected back toward the source. This results in a difference in intensity between the two ears. As seen in Figure 2.2, IIDs are insignificant at low frequencies but are as great as 20 dB at greater frequencies.

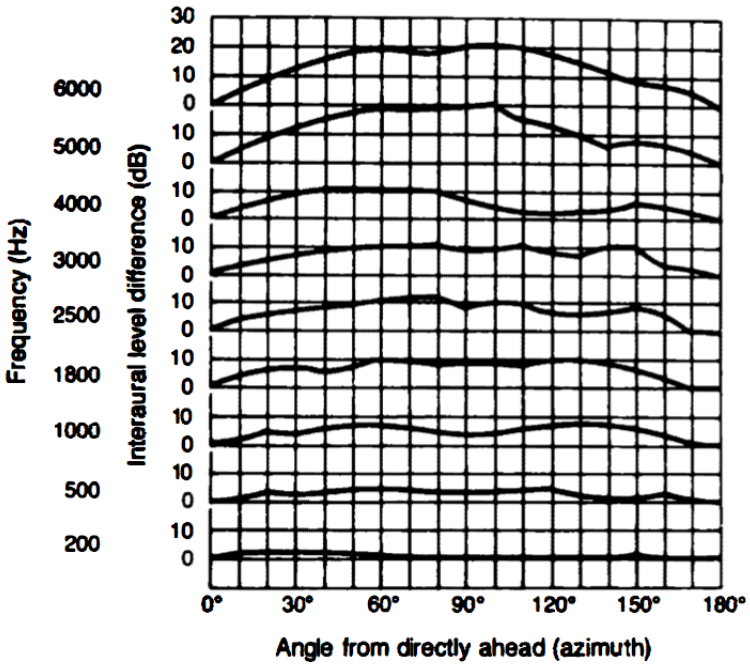


Figure 2.2: Figure showing IID as a function of azimuth from Feddersen *et al.* The stimulus in this case in sinusoidal and each curve corresponds to a different frequency.

2.2 The precedence effect

In reverberant environments, a sound that is produced propagates in many directions and is reflected from nearby surfaces in its path. While the task of resolving the direct sound from all reflections is a difficult one, the human auditory system manages to localize sounds quite robustly even through this clutter of information.

For example, consider an arrangement of two loudspeakers in an anechoic room such that the speakers are at the same distance from the listener, and stimulated by identical sounds such that the onset of one sound is delayed relative to the onset of the other sound. This is illustrated in Figure 2.3. This can be considered a model of a direct sound with a single reflection. If two sounds arrive at the ears with a short delay of less than 20 ms or so between them, the two sounds appear perceptually fused. It has also been seen that the location of the fused sound appears to be dominated by the location of the lead

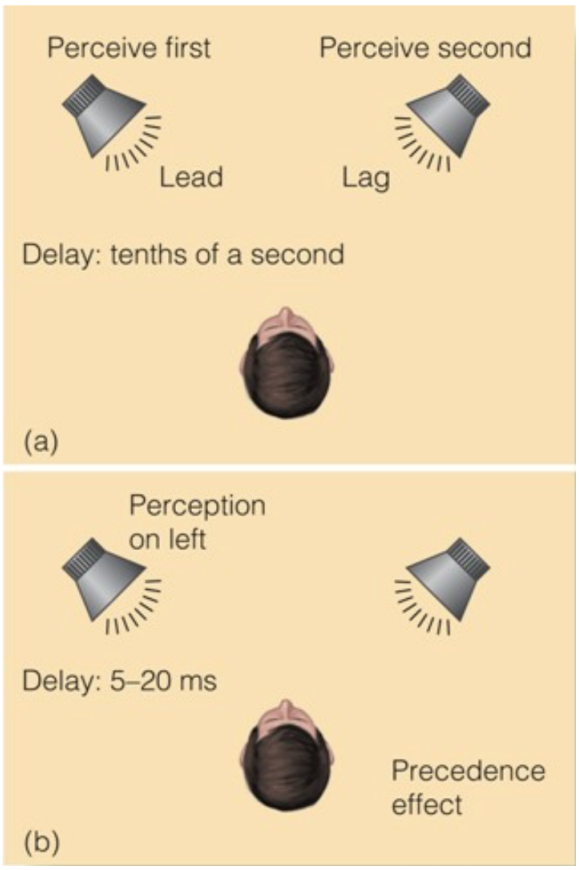


Figure 2.3: Figure illustrating the precedence effect. Two sounds are played with a relative delay with respect to each other a) Delay of tenths of a second leading to perception of two separate audio events b) Delay of 5-20 ms leading to localization that is dominated by the lead sound

sound i.e. the sound that was played first. This is called the "law of the first wave front" or the "precedence effect". As the delay increases beyond a certain threshold, the two sounds become two audibly separate events. The precedence effect is considered to be a major contributing factor behind robust auditory perception by humans in the presence of reverberation.

2.3 Models of binaural hearing

In this section, we will review a few important models of binaural interaction. Among the most seminal theories were the coincidence-based model proposed by Jeffress [8] and the equalization-cancellation model proposed by Durlach [9]. Most modern theories of binaural interaction have their roots in the two models mentioned above. Colburn's quantification of the Jeffress hypothesis is also discussed briefly [10].

2.3.1 Coincidence-based model

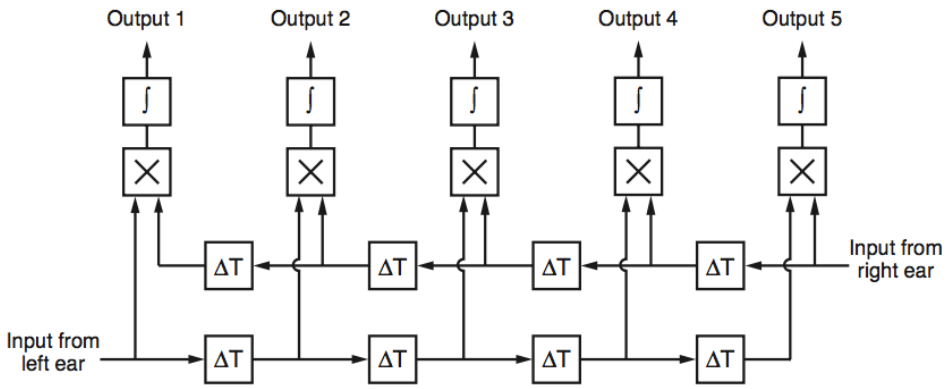


Figure 2.4: Schematic diagram of Jeffress place mechanism from [11]. Boxes labelled with crosses are multipliers that record coincidences between the neural activity of the two ears.

Jeffress postulated a mechanism that consisted of a number of central neural units that recorded coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same Characteristic Frequency (CF). He further postulated that the neural signal coming from one of the two fibers is delayed by a small amount that is fixed for a given fiber pair, as seen in Figure 2.4. Because of the synchrony in the response of low-frequency fibers to low-frequency stimuli, a given binaural coincidence-counting unit

at a particular frequency will produce maximal output when the external stimulus ITD at that frequency is exactly compensated for by the internal delay of a fiber pair. Hence, the external ITD of a simple stimulus could be inferred by determining the internal delay that provides the greatest response over a range of frequencies. While the delay mechanism was conceptualized by Jeffress and others in the form of the ladder-type delay, such a structure is only one of several possible realizations. The important characteristic-delay parameter of the ITD-sensitive units is represented by the difference in total delay incurred by the neural signals from the left and right ears that are input to a particular coincidence-counting unit. The short-term average of a set of such coincidence outputs at a particular CF plotted as a function of their internal delay is an approximation to the short-term cross-correlation functions of the neural signals arriving at the coincidence detectors.

2.3.2 The Equalization-Cancellation model

The Equalization-Cancellation (EC) model was first suggested by Kock [12] and was subsequently developed extensively by Durlach [9]. The EC model assumes that the auditory system transforms the signals arriving at the two ears so that the masker components are "equalized" by imposing a delay and an amplitude change to the signal on one side, with the goal of making the two signals equal to one another to the extent possible. Detection of the target is achieved by "cancelling", or subtracting the signals to the two ears after the equalization operation. Quantitative predictions for the EC model are obtained by specifying limits to the operations used to achieve the cancellation process, as well as sources of internal noise.

2.3.3 Jeffress-Colburn model

Colburn [10] reformulated the Jeffress hypothesis quantitatively using a relatively simple model of the auditory-nerve response to sound as Poisson processes, and a "binaural displayer" consisting of a matrix of coincidence-counting units of the type postulated by Jeffress. These units are specified by the CF of the auditory-nerve fibers that they receive input from as well as their intrinsic internal delay. The overall response of an ensemble of such units as a function of internal delay is a representation that is similar to the running interaural cross-correlation of the signals to the two ears, after the peripheral cochlear analysis leading to the representation at the level of the auditory-nerve fibers [13].

2.4 Automatic Speech Recognition basics

Automatic Speech Recognition (ASR) refers to the process of using a computer to automatically transcribe spoken words into text. The first speech recognition systems were only able to recognize words spoken in isolation by a known speaker based on a concept called dynamic time warping (DTW). These systems had models for entire words meaning that every word in the ASR vocabulary needed to be known. Transcription was performed by finding the word closest to the one that was spoken among all models.

The development of the Hidden Markov Model (HMM) significantly improved ASR accuracy. An HMM is a first-order probabilistic characterization of a time-varying process that allows for a maximum likelihood prediction of a system of state sequences given a series of observations. In ASR, the series of observations is the sampled speech signal, and the states are either words or some atomic unit of words (e.g., phonemes). Since all words can now be represented by a small set of atomic sound units, this allows for huge

vocabularies to be modeled efficiently.

2.4.1 Feature computation

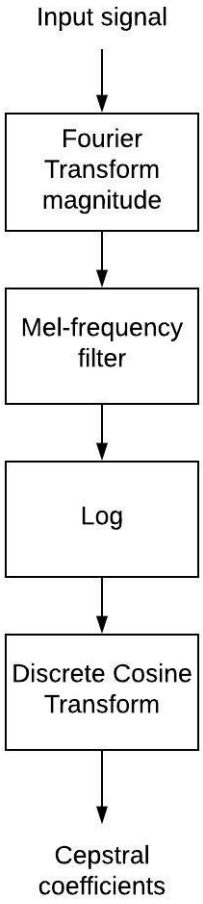


Figure 2.5: Block diagram for MFCC processing.

The raw speech signal does not effectively present the information that is relevant to speech recognition. For this reason, feature extraction methods are used to transform the raw signal into a feature vector. This effectively provides a more compact and relevant representation of the original signal. The most widely used feature set is the Mel-Frequency Cepstral Coefficient (MFCC) features. As mentioned above, the use of MFCC features re-

duces the dimensionality of the data being fed to the ASR engine. MFCC features also reduce the variability across speakers and noise conditions. Figure 2.5 is a block diagram showing the different stages of MFCC processing.

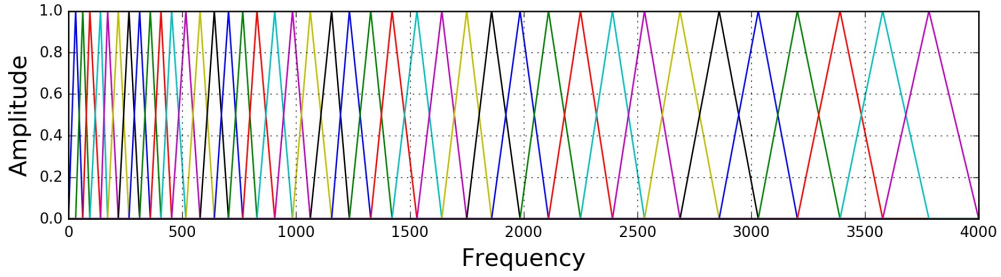


Figure 2.6: Mel-scale weighting functions.

The first step involves taking the short-time Fourier transform magnitude of the input signal, typically using a Hamming window of 20-35 ms in duration. MFCC processing achieves smoothing and dimensionality reduction by applying a set of frequency-selective weighting functions to the magnitude spectrum. The weighting functions, as shown in Figure 2.6, are typically triangular in shape and are spaced according to the perceptually-motivated Mel scale. Each feature dimension is computed as the dot product of each triangular weighting function with the Fourier transform magnitude, which effectively accomplishes a form of bandpass filtering of the signals in each channel.

The dynamic range of a typical speech spectrum often spans several orders of magnitude. As seen in Figure 2.5, a logarithmic non-linearity is applied to the signal post bandpass-filtering. This shrinks the dynamic range of the observed spectrum, allowing small deviations to be more easily captured by the acoustical model.

The final step in MFCC processing is the application of the Discrete Cosine Transform (DCT), which can be thought of as a Fourier series expansion of the log of the magnitude of the spectrum. The first 12 or so coefficients describe the shape of the vocal tract filter which mediates the production of an ensemble of perceptually-distinct phonemes, while the higher coefficients indicate the fundamental frequency of voicing. Feature values for ASR are typically truncated to 12-14 cepstral coefficients.

The DCT exhibits an energy compaction property such that the energy of its coefficients are highly concentrated at low indices. Consequently, truncating the DCT causes relatively little information in the signal to be lost.

2.4.2 Traditional ASR

Speech recognition is a special case of the more general Bayesian classification problem. Given a sequence of observations X, the ASR engine determines the most likely sequence of phonemes (or words) $\hat{W}$

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\mathbf{argmax}}\, \boldsymbol{Pr}(\boldsymbol{W}|\boldsymbol{X}) \tag{2.1}$$

Applying Bayes rule to Equation 2.1 reveals the two primary system components of a speech recognition system, the acoustic model, and the language model,

$$\begin{aligned} \hat{W} &= \underset{W}{\operatorname{argmax}} \frac{Pr(X|W)Pr(W)}{Pr(X)} \\ &= \underset{W}{\operatorname{argmax}} \underbrace{Pr(X|W)}_{\text{Acoustical model}} \underbrace{Pr(W)}_{\text{Language model}} \end{aligned} \tag{2.2}$$

Using a large number of example utterances from a given language, the language model (LM) characterizes the probability of observing a given sequence of words in that language. LMs are based on the notion of an n-gram, which models the probability of a word or phoneme given the previous $n-1$ words or phonemes. Usually, bigrams or trigrams are used.

The acoustical model, on the other hand, characterizes the probability of observing a particular manifestation of a speech sound in the feature space. In an ASR engine, the audio stream is broken up into overlapping frames. Each of these frames is transformed into a set of cepstral coefficients as described in Section 2.4.1. For different manifestations

of a particular speech sound, clusters of cepstral coefficients are seen and it is this that the acoustical model captures. Acoustical modeling allows the ASR system to make a probabilistically-optimal decision as to what sound is most likely being made.

The HMM representation characterizes the incoming speech waveform as a doubly stochastic process [14]. First, the sequence of phonemes that are produced is characterized as a set of unobserved Markov states which presumably represent the various configurations that the speech production mechanisms may take on. As is the case for all Markov models, the transition probabilities depend only on the current state that is being occupied. Each state transition causes a feature vector to be emitted that is observable, with the probability density of the components of the feature vector depending on the identity of the state transition. The task of the decoder is to infer the identity of the unobserved state transitions. Gaussian mixture densities are commonly used for the phonetic models, in part because the parameters of these densities can be estimated efficiently. HMMs using gaussian mixtures for the phonetic models are frequently referred to as "HMM-GMM" systems.

2.4.3 Deep Neural Networks for acoustical models

As mentioned earlier, HMM-GMM systems have been quite successful in the past in terms of ASR performance. However, GMMs are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. On the other hand, neural networks trained by backpropagating error derivatives can learn this sort of data much better. The resources needed to train a complex neural net may have made it difficult to try in the past. However, with the current advances in hardware, better training methodologies, and (more than anything) the availability of large training databases, Deep Neural Networks (DNNs) are becoming increasingly effective as a replacement for GMMs.

A two-stage training procedure is typically used for training the DNNs. In the first

stage, a stack of generative models, each with one layer of latent variables, is used to initialize feature detectors one layer at a time. Restricted Bolzmann machines (RBMs), which are a type of Markov Random Field, are usually used for this purpose. In the second stage, each generative model is used to initialize one layer of hidden units in a DNN and then the network is discriminatively tuned to predict target HMM states. The targets are obtained using forced alignment with a baseline HMM-GMM system.

2.5 Brief description of relevant algorithms

This section briefly goes over two algorithms that have been extensively used in this study. The Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) algorithm performs steady-state suppression and has been shown to be very effective in reverberant conditions. The Phase Difference Channel Weighting (PDCW) algorithm helps with isolating a target talker in the presence of one or more interfering talkers using ITD-based analysis done in the frequency domain.

2.5.1 Suppression of Slowly-varying components and the Falling edge of the power envelope

The Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) algorithm [4, 15] was used in this study to achieve steady-state suppression. The SSF algorithm is motivated by the precedence effect and by the modulation-frequency characteristics of the human auditory system. A block diagram describing SSF processing is shown in Figure 2.7. SSF processing is performed separately on each channel of the binaural signal.

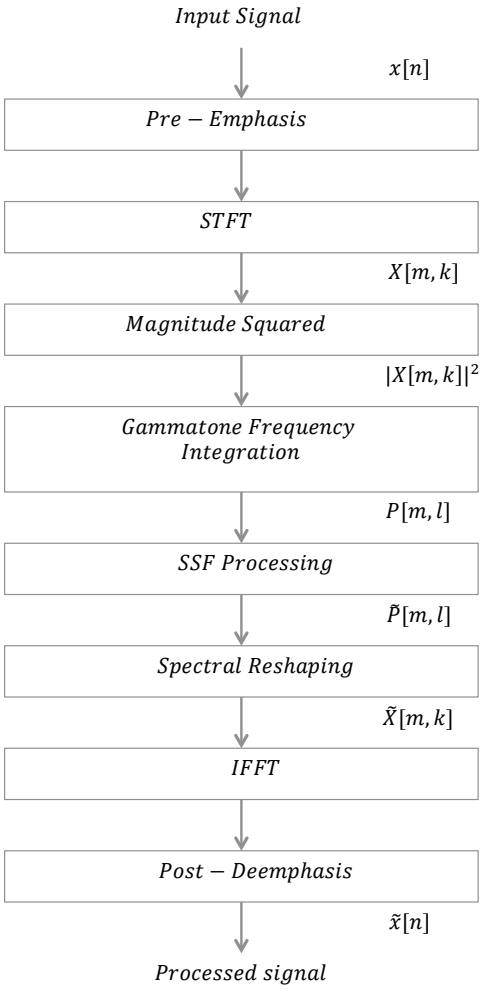


Figure 2.7: Block diagram describing the SSF algorithm.

After performing pre-emphasis on the input signal, a Short-Time Fourier Transform (STFT) of the signal is computed using a 40-channel gammatone filterbank. The center frequencies of the gammatone filterbank are linearly spaced in Equivalent Rectangular Bandwidth (ERB) [16] between 200 Hz and 8 kHz. In general, longer-duration window sizes for STFT computation have been shown to be useful for noise compensation [17, 4]. The power $P[m, l]$ corresponding to the m^{th} frame and the l^{th} gammatone channel is given by,

$$P[m, l] = \sum_{k=0}^{N-1} |X[m, k]H_l[k]|^2, 0 \leq l \leq L - 1, \tag{2.3}$$

where $H_l[k]$ is the frequency response of the l^{th} gammatone channel evaluated at the k^{th}

frequency index and $X[m,k]$ is the signal spectrum at the m^{th} frame and the k^{th} frequency index. N is FFT size which was 1024.

The power $P[m,l]$ is then lowpass filtered to obtain $M[m,l]$.

$$M[m,l] = \lambda M[m-1,l] + (1-\lambda)P[m,l], \tag{2.4}$$

Here λ is a forgetting factor that was adjusted for the bandwidth of the filter and experimentally set to 0.4. Since SSF is designed to suppress the slowly-varying portions of the power envelopes, the SSF processed power $\tilde{P}[m,l]$ is given by,

$$\tilde{P}[m,l] = max(P[m,l] - M[m,l], c_0 M[m,l]), \tag{2.5}$$

where c_0 is a constant introduced to reduce spectral distortion. Since $\tilde{P}[m,l]$ is obtained by subtracting the slowly varying power envelope from the original power signal, it is essentially a highpass-filtered version of $P[m,l]$, thus achieving steady-state suppression. The value for c_0 was experimentally set to 0.01.

For every frame in every gammatone filter band, a channel-weighting coefficient $w[m,l]$ is obtained by taking the ratio of the highpass filtered portion of $P[m,l]$ to the original quantity.

$$w[m,l] = \frac{\tilde{P}[m,l]}{P[m,l]}, 0 \leq l \leq L-1 \tag{2.6}$$

Each channel-weighting coefficient corresponding to the l^{th} gammatone channel is associated with the response $H_l[k]$ and so the spectral-weighting coefficient $\mu[m,k]$ is given by

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l]|H_l[k]|}{\sum_{l=0}^{L-1} |H_l[k]|}, 0 \leq l \leq L-1, 0 \leq k \leq N/2 \tag{2.7}$$

The final processed spectrum is then given as

$$\tilde{X}[m,k] = \mu[m,k]X[m,k], 0 \leq k \leq N/2 \tag{2.8}$$

Using Hermitian symmetry, the rest of the frequency components are obtained and the processed speech signal $\tilde{x}[n]$ is re-synthesized using the overlap-add method.

The SSF algorithm is very effective in mitigating the effect of reverberation.

2.5.2 Phase Difference Channel Weighting

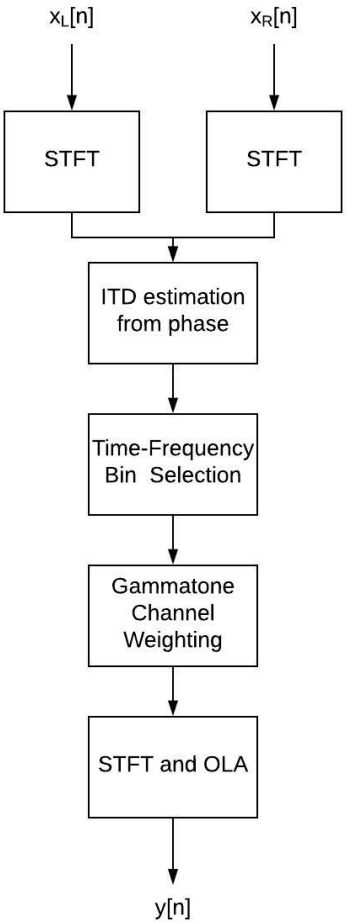


Figure 2.8: Block diagram describing the PDCW algorithm.

The Phase Difference Channel Weighting (PDCW) algorithm separates signals according to ITD, in a crude approximation to human sound separation. PDCW estimates ITD indirectly, computing interaural phase difference (IPD) information in the frequency domain

and then dividing by frequency to produce estimated ITDs. Again, it is assumed that there is no delay in the arrival of the target signal between the right and left channel.

A block diagram detailing the PDCW algorithm is seen in Figure 2.8. The PDCW algorithm starts with applying a Short-Time Fourier Transform (STFT) to the input signals from the left and right microphones $x_L[n]$ and $x_R[n]$. Since the two signals are identical except for a time delay, the phase difference between signals from the two microphones is calculated using the STFT phase. The frequency-dependent ITD $d(k,m)$ for time-frequency bin (k,m) is given by,

$$|d(k,m)| \approx \frac{1}{|\omega_k|} \min_r |\angle X_R(k,m) - \angle X_L(k,m) - 2\pi r| \tag{2.9}$$

where $X_R(k,m)$ and $X_L(k,m)$ are the STFT of $x_R[n]$ and $x_L[n]$ respectively.

Components of the STFT are retained if they are within zero ITD by a threshold amount τ in magnitude. A binary mask $\mu(k,m)$ is derived for the k^{th} time frame and the m^{th} frequency channel using the ITD $d(k,m)$ such that, $\mu(k,m) = 1$ for components with ITD less than the threshold magnitude and some very small quantity η otherwise.

$$\begin{aligned} \mu(k,m) &= 1, |d(k,m)| \leq \tau \\ &= \eta, otherwise \end{aligned} \tag{2.10}$$

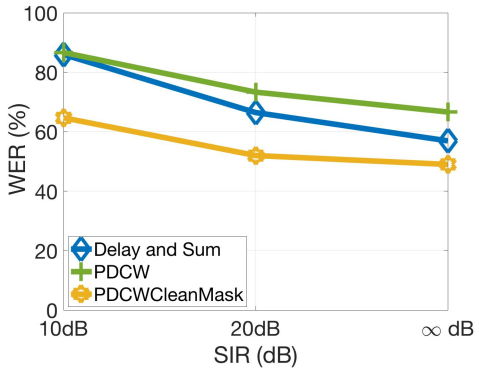
While this mask provides a degree of signal separation by itself, recognition accuracy improves when it is smoothed over frequency. This smoothing along frequency, called "channel weighting" in the original algorithm, is performed using a gammatone weighting function. PDCW provides substantial improvements in ASR accuracy in the presence of interfering talkers, although its performance degrades sharply in the presence of reverberation [6].

Effect of reverberation on PDCW

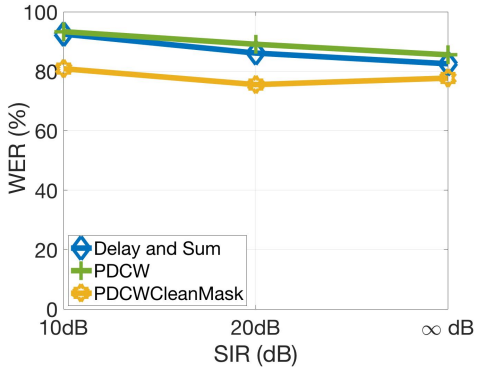
As mentioned above, the effectiveness of the PDCW algorithm sharply declines in the presence of reverberation. There could be a number of reasons for the poor performance of PDCW in the presence of reverberation. It could be that due to the room modes in the presence of reverberation, the acoustic target might no longer be in the location of the actual target. It might be necessary to find target location first for ITD-based analysis. On the other hand, it could be that the presence of reverberation causes the ITD estimation to be noisy. In order to investigate further, a short experiment was conducted.

PDCW was applied as described in Section 2.5.2 to speech signals for ASR experiments. However, mask estimation was done using the clean version of the speech signals which had no reverberation. The interfering talker was still present however. The hypothesis was that if using the mask estimates from clean speech lead to an improvement in ASR performance, that would mean that the location of the acoustic target is the same as the original target location. That would mean that noisy ITD estimates would most likely be the reason for poor ASR performance. The results obtained are shown in Figure 2.9.

Figure 2.9 compares PDCW using speech in a reverberant environment to PDCW run with masks estimated from speech signals that had no reverberation (referred to as PDCWCleanMask in the figure). The ASR experiments were run using the RM1 database and the Sphinx speech recognition engine. Results for the Delay and Sum algorithm have also been plotted as a baseline. As seen, the use of masks derived from speech without reverberation does lead to significant improvements in ASR performance. PDCW using reverberant speech for mask estimation has Word Error Rate (WER) higher than the baseline Delay and Sum algorithm consistently. However, the use of clean speech-based masks alters this significantly especially at Signal-to-Interference Ratios (SIRs) of 10 and 20 dB SIR. Due to the reverberation that is present, the WER is still considerably high even with PDCWCleanMask. But the drop in error is significant enough to conclude that the acous-



(a)



(b)

Figure 2.9: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the RM1 database and the Sphinx speech recognition engine: (a) 0.5 s (b) 1 s. The PDCW algorithm, using reverberant speech for mask estimation is compared to the use of speech with no reverberation for mask estimation.

tic target is the same and it is the ITD estimation that is causing the very high error rates. The presence of reverberation produces reflections that are added to the direct response in a fashion that leads to unpredictable phase changes, which essentially makes the ITD estimation much less accurate. Further details about the algorithm are provided in [18].

CHAPTER 3

EXPERIMENTAL CONFIGURATION

The primary goal of this work is to exploit techniques based on human binaural processing to improve recognition of speech in complex acoustic environments, specifically in the presence of multiple talkers and reverberation. To this end, a reverberant acoustic environment that has interfering talkers present has been simulated for this work as described in Section 3.2. A definition of the problem that we are attempting to solve is given in Section 3.1. Section 3.3 describes how the performance of the various algorithms discussed in this thesis are evaluated using ASR experiments.

3.1 Problem definition

Figure 3.1 depicts the microphone configuration used in this study. As seen, we consider specifically a two-microphone configuration. These microphones are assumed to be placed somewhere in a reverberant room. There is a target source present that is in front of the two microphones in such a way that that the speech from the target arrives at the two microphones simultaneously. In addition, an interfering source is also present that is located at some angle ϕ with respect to the target source. In general, the problem can be

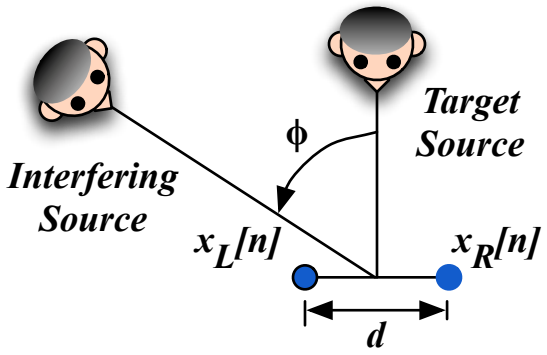


Figure 3.1: Two-microphone configuration used in this study with an on-axis target source and an off-axis interfering source.

expressed as separation of a target signal from all the other signals present in the given acoustic environment, including both speech from the interfering source and reverberant target components arising from reflections by the surfaces of and objects in the room.

3.2 Simulated Data

Most of the experiments performed in this study used simulated data. Data were simulated in accordance with the setup described in Section 3.1. A reverberant rectangular room was simulated using the well known "Image Method" [19]. A room of dimensions $5m \times 4m \times 3m$ was assumed. The distance between the two microphones is 4 cm. The target speaker is located $2m$ away from the microphones along the perpendicular bisector of the line connecting the two microphones. An interfering speaker is located at an angle of 45 degrees to one side and $2m$ away from the microphones. The sources and microphones are all $1.1m$ above the floor. To simulate the interfering talker, speech from the same database as the target speech was used. The interfering speech signal was mixed in at various levels of Signal-to-Interference Ratio (SIR).

3.3 Performance evaluation

3.3.1 Speech recognition engine

ASR experiments were conducted using the CMU SPHINX-III speech recognition system and the Kaldi speech recognition toolkit. The DARPA Resource Management (RM1) and Wall Street Journal (WSJ) databases [20] were used. The training set for RM1 consisted of 1600 utterances and the test set consisted of 600 utterances. For WSJ, these numbers were 7138 and 330 respectively. Features used were 13^{th} order mel-frequency cepstral coefficients.

The Kaldi speech recognition toolkit allows for DNN-based acoustical models to be trained. Thus, preliminary ASR results were obtained using Sphinx and the RM1 database. The reverberation times tested for the preliminary results were 0.5 s and 1 s. For each reverberation time, an interfering talker was mixed in at 0, 10 and 20 dB SIR. Experiments were also conducted in the absence of an interferer. More detailed experiments were then performed using GMM-based and DNN-based acoustical models trained using Kaldi. Clean as well as multi-style training was used in most cases. These experiments used simulated data using the WSJ database at reverberation times of 0.2 s to 1 s in steps of 0.2 s with an interfering talker mixed in at 10 and 20 dB SIR. As before, experiments were also conducted in the absence of an interferer.

3.3.2 Performance metric

The metric used to evaluate the effectiveness of the algorithms discussed in this thesis is Word Error Rate (WER) in ASR experiments. WER is defined as the ratio of total insertions,

deletions, and substitutions in the transcript generated by the ASR engine to the number of words spoken by the target source.

$$WER = \frac{(S + D + I)}{N} \tag{3.1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the total number of words in the reference transcript. As seen in Equation 3.1, WER can be greater than a hundred percent.

3.4 Effect of locations of microphones in the room

An observation made early on was that the ASR performance was significantly impacted by the location of the microphones in the room. Because of the manner in which reverberation is simulated, it is possible that there are regions of constructive or destructive interference in the room. These regions might therefore cause the signal to be considerably altered leading to differences in ASR performance based on location alone. In order to test the efficacy of the algorithm under test independent of location or any other external factors, data were simulated such that several different room locations were utilized for the microphones.

To determine how many microphone locations were necessary in order to ensure that the room location does not play a major role in determining the performance of an algorithm under test, an experiment was conducted. Test data were generated using M different locations for each trial and ASR experiments were conducted. For each value of M, there were N trials to see how much the WER changed for each trial using the same number of locations. It is to be noted that the entire setup as described in Figure 3.1 was moved with the relative positions of the target, interferer and microphones kept unchanged.

Experiments were conducted using the Delay-and-Sum algorithm. The Sphinx speech recognition engine was used for the experiments. The database used for these experiments was RM1. The value of M used varied between 5-30 locations in steps of 5. A total of 10 trials were run for each value of M. The aim was to find out the value of M for which the WER across all the trials remains reasonably similar. The results obtained are shown in Figure 3.4. The WER is plotted in Figure 3.4 as a function of the variable M. The mean WER for each value of M as well as the standard error of the mean is also shown.

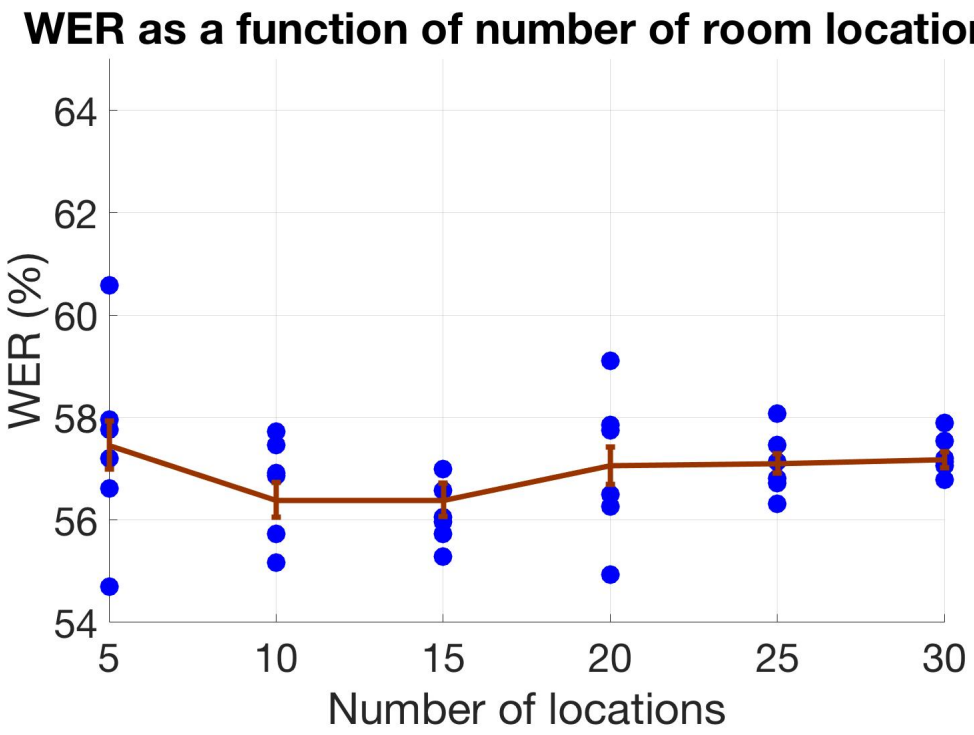


Figure 3.2: Word Error Rate as a function of the number of room locations used for testing the Delay-and-Sum algorithm. A total of 10 trials were run for a given number of room locations used. The average WER as well as the standard error of the mean is also shown.

As expected, when fewer room locations were sampled, the WER varied quite a bit from trial to trial. The standard deviation of the WER across the 10 trials performed keeps decreasing as the number of locations is increased. The value of $M = 25$ was selected based on these results.

CHAPTER 4

ITD-BASED BINAURAL PROCESSING USING SIGNAL ENVELOPES

In this chapter we introduce and discuss an algorithmic approach that is based on the concept of exploiting correlations across the sensors (which in this case consists of two microphones). Signal components originating from the same source tend to be mutually coherent. As seen in Figure 2.1, the sound emanating from Source A arrives at both ears at the same time thus leading to an ITD of 0. However, sounds emanating from Source B will have a non-zero ITD that can be computed. In the specific experimental setup described in Figure 3.1, signals originating from the target source will ideally be perfectly coherent across the two microphones, with ITD equal to 0. On the other hand, the interferer that is off to one side will arrive at the left microphone before the right microphone leading to an inter-microphone delay time. This delay between the signals can be computed. For the target signal given by $r(t)$ and an interferer signal given by $i(t)$ as shown in Figure 4.1, the interferer $i(t)$ travels extra distance to arrive at the right ear. If the signal from the interferer arrives with an azimuth angle of θ radians and the distance between the two sensors is d meters, this extra distance traveled is $dsin\theta$. Thus, the time delay τ between the two sensors is,

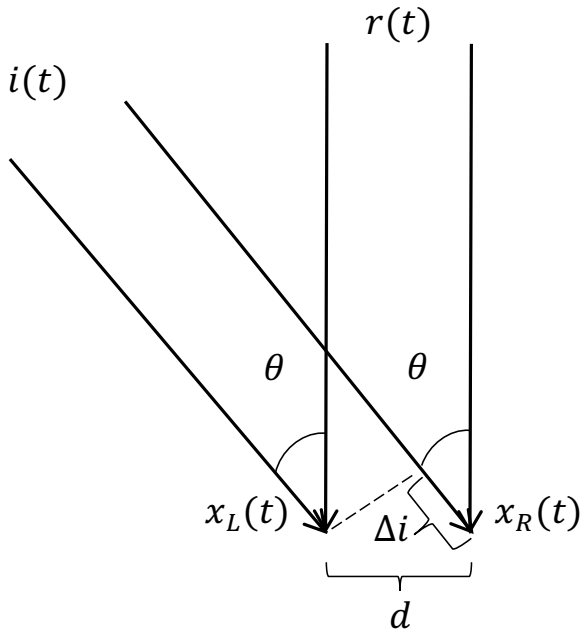


Figure 4.1: Experimental setup used in this study.

$$\tau = \frac{d \sin \theta}{c} \tag{4.1}$$

where c is the speed of sound.

As discussed in Chapter 2.1, the reliability of ITD cues is dependent on frequency. At low frequencies, ITD-based cues reliably predict the location of a sound source. This is not the case at higher frequencies, typically greater than 1500 Hz. At higher frequencies, amplitude envelopes are likely to be the cue that is utilized for the computation of ITD-based cues for localization. The importance of amplitude envelopes was discussed in [21], among other sources. This study found that the detectability of ITDs in the envelope of a high frequency carrier was about the same as a pure tone with the frequency of the envelope. Among human subjects, it was the ITD of the envelope rather than that of the fine structure that determined how subjects performed in the lateralization tasks.

Several models of binaural interaction in the human auditory system are discussed in Chapter 2. Several of these models are based on the cross-correlation of the signals to the two ears typically preceded by processing in the auditory periphery. This mechanism has a physiological correlate as seen by the presence of cells in the superior olivary complex and the inferior colliculus of the brainstem (*e.g.* [22]) that are sensitive to signals presented with a specific ITD. In this chapter, we present a new method that uses a simple interaural cross-correlation-based weighting using speech envelopes to isolate the target signal based on ITD [23]. ITD is computed using the interaural (or, more accurately, inter-microphone) cross-correlation of the signals. As mentioned in Chapter 3, the specific experimental conditions addressed in this work are a combination of noise and reverberation. ITD estimation is especially compromised in the presence of reverberation. The reverberant field blurs the ITD estimates significantly making localization much more difficult in reverberant environments.

The "precedence effect" ([24, 25, 26]), as discussed in Chapter 2, is considered to be one of the important mechanisms mediating human auditory perception in the presence of reverberation. The precedence effect describes the phenomenon by which directional cues due to the first-arriving wavefront (corresponding to the direct sound), are given greater perceptual weighting than those cues that arise as a consequence of subsequent reflected sounds. The precedence effect is thought to have an underlying inhibitory mechanism that suppresses echoes at either the monaural level [27] or binaural level [28]. Considering the monaural approach, a reasonable way to overcome the effects of reverberation would be to boost these initial wavefronts. This can also be achieved by suppressing the steady state components of a signal.

The algorithm described in this chapter presents a combination of the concepts of precedence-effect-based processing and ITD analysis to improve recognition accuracy in environments containing reverberation and interfering talkers. A novel method for ITD analysis is introduced, called the Interaural Cross-correlation-based Weighting (ICW) al-

gorithm. In addition, cross-correlation-based weighting using the ICW algorithm is preceded by suppression of reverberant components of speech using steady state suppression. The SSF algorithm proposed by C. Kim ([4, 15]) as described in Section 2.5.1 was used to achieve this.

4.1 Motivation based on auditory processing

Interaural cross-correlation is commonly used in binaural processing to compute the delay between the signal at the two ears, producing an estimate of the ITD. The ICW algorithm uses this principle for ITD-based processing. The ICW algorithm roughly follows the manner in which binaural signals are processed by the human auditory system.

A crude model of the auditory-nerve response to sounds starts with bandpass filtering of the input signal (modeling the frequency selectivity of the cochlea), followed by half-wave rectification and then by a lowpass filter. The auditory-nerve response roughly follows the fine structure of the signal at low frequencies and the envelope of the signal at high frequencies [3, 11, 13]. ITD analysis is based on the cross-correlation of auditory-nerve responses. Hence, the human auditory system is especially sensitive to fine-structure ITD cues at low frequencies and envelope ITD cues at high frequencies. The ICW algorithm uses this concept to reject components of the input signal that appear to produce large envelope ITDs that are unlikely to represent target components. Moreover, signal envelopes have been shown to be a better measure in detecting the direct sound [27, 29].

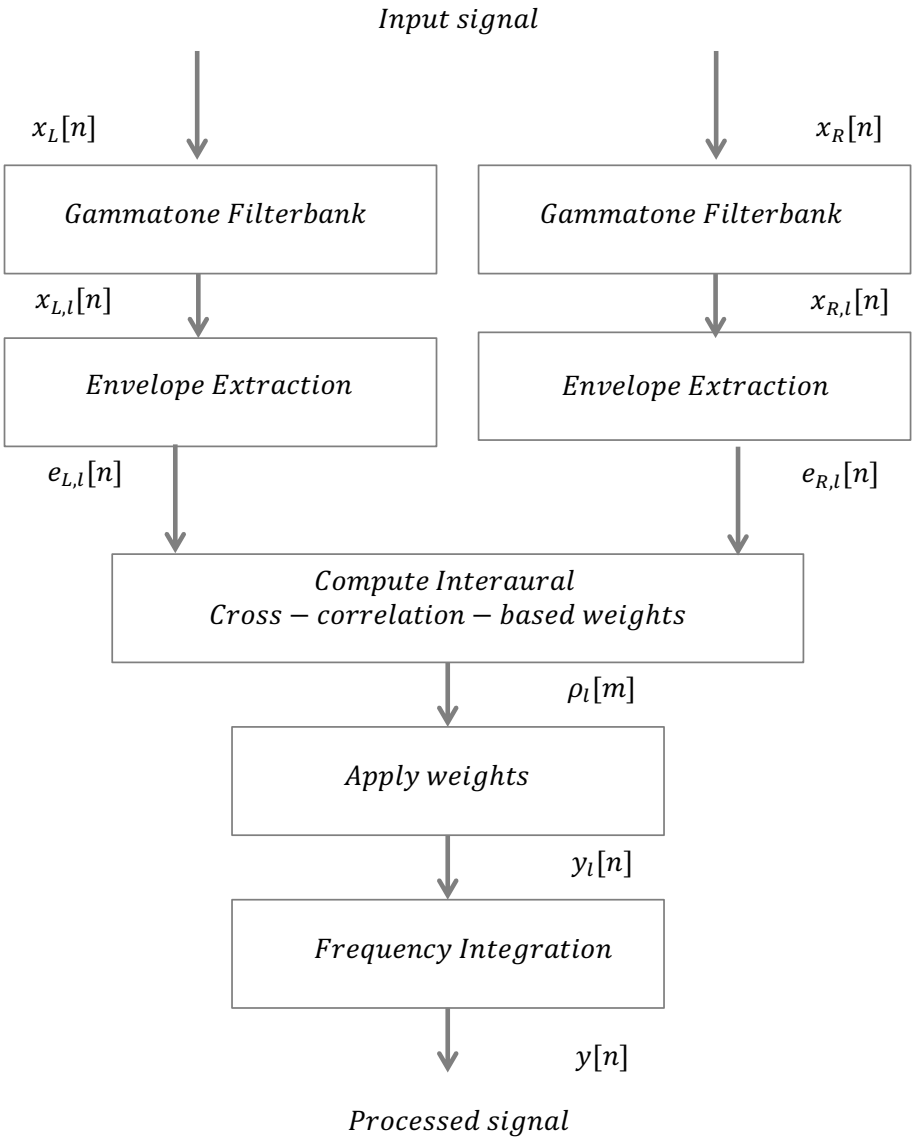


Figure 4.2: Block diagram describing the ICW algorithm.

4.2 Structure of the ICW algorithm

Figure 4.2 shows a block diagram of the ICW algorithm. As mentioned in Section 3.1, it is assumed that there is no delay in the arrival of the target signal between the right and left channel denoted by $x_R[n]$ and $x_L[n]$ respectively.

4.2.1 Bandpass Filtering

The signals $x_R[n]$ and $x_L[n]$ are first bandpass filtered by a bank of 40 gammatone filters using a modified version of the implementation in Malcolm Slaney's Auditory Toolbox [30]. The center frequencies of the filters are linearly spaced according to their equivalent rectangular bandwidth (ERB) [16] between 100 Hz and 8 kHz. The use of gammatone filters is physiologically motivated, as they mimic the form of frequency analysis in the peripheral auditory system. The frequency response of gammatone filters is shown in Figure 4.3

In order to ensure that there is no delay across the different frequency channels, zero-phase filtering was performed. A zero-phase filter has no phase distortion and for a real impulse response $h(n)$, it satisfies,

$$h(n) = h(-n)$$

Zero-phase filtering was achieved using forward-backward filtering. A signal is first filtered normally using an impulse response $h(n)$. The output of the first filtering operation is then filtered again using the flipped version of the original filter i.e. $h(-n)$. Thus the effective filtering operation has the impulse response of,

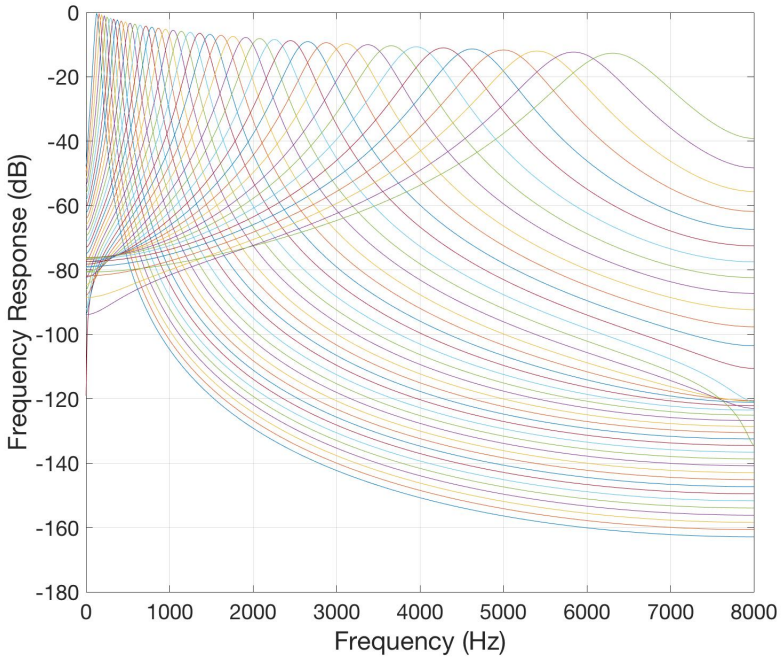


Figure 4.3: Frequency response of gammatone filters used in this study. This was generated using Malcolm Slaney's Auditory Toolbox.

$$h_e(n) = h(n) * h(-n) \tag{4.2}$$

As seen this effective filter h_e is zero phase. In the frequency domain,

$$\begin{aligned} H_e(e^{j\omega t}) &= H(e^{-j\omega t})H^*(e^{-j\omega t}) \\ &= |H(e^{j\omega t})|^2 \end{aligned} \tag{4.3}$$

Thus, effectively, $H_e(e^{j\omega t})$ squares the amplitude response and zeros the phase response of the original filter $H(e^{j\omega t})$. Since this leads to an effective reduction in bandwidth, the bandwidths of the original gammatone filters are modified to roughly compensate for this.

4.2.2 Envelope extraction

After the peripheral bandpass filtering, the envelopes $e_{L,l}[n]$ and $e_{R,l}[n]$ of the signals are extracted. Here, $e_{L,l}[n]$ refers to the envelope of the signal from the left microphone and the l^{th} gammatone filter channel and $e_{R,l}[n]$ is the envelope of the signal from the right microphone and the l^{th} gammatone filter channel. The Hilbert transform was used for envelope extraction.

The analytic signal corresponding to the left microphone signal $x_{L,l}[n]$ is given by,

$$x^a_{L,l}[n] = x_{L,l}[n] + j\mathcal{H}(x_{L,l}[n]) \tag{4.4}$$

where $\mathcal{H}(x_{L,l}[n])$ is the Hilbert transform of $x_{L,l}[n]$. The analytic signal is a complex valued signal with no negative frequency components. The real and complex parts of the analytic signal are related by the Hilbert transform. The instantaneous envelope of the signal $x_{L,l}[n]$ can be computed using the analytic signal.

$$\begin{aligned} e_{L,l}[n] &= |x^a_{L,l}[n]| \\ &= |x_{L,l}[n] + j\mathcal{H}(x_{L,l}[n])| \end{aligned} \tag{4.5}$$

The Hilbert transform basically introduces a phase shift of 90° in the original signal.

The microphone signal on the right $x_{R,l}[n]$ is also processed in exactly the same manner to give the envelope $e_{R,l}[n]$.

4.2.3 Cross-correlation and computation of a weight matrix

Once the signal envelopes are computed, they are divided into frames by windowing. Thereafter, cross-correlation across the right and left microphone signals is performed.

The normalized cross-correlation of the envelope signals $e_{R,l}[n]$ and $e_{L,l}[n]$ is given by,

$$\rho_l[m] = \frac{\sum_{N_w} e_{L,l}[n;m] e_{R,l}[n;m]}{\sqrt{\sum_{N_w} e_{L,l}[n;m]^2} \sqrt{\sum_{N_w} e_{R,l}[n;m]^2}} \tag{4.6}$$

where $\rho_l[m]$ refers to the normalized cross-correlation of the m^{th} frame and l^{th} gammatone channel, $e_{L,l}[n;m]$ and $e_{R,l}[n;m]$ are the envelope signals corresponding to the m^{th} frame and l^{th} gammatone channel for the left and right channels respectively. The window size N_w was set to 75 ms and the time between frames for ICW was 10 ms. Rectangular windows were used.

As discussed in Chapter 3, the microphone arrangement is such that the target signal produces zero delay between the right and left microphones. For this reason, normalized cross-correlation was performed only for a delay of zero corresponding to the target delay. This cross-correlation computation would lead to a $\rho_l[m] = 1$ if the two signals are identical, as would be the case ideally. In the presence of noise and reverberation, the value of $\rho_l[m]$ is indicative of the degree of mismatch between the right and left channels. The differences between the two signals are caused due to the reverberated signal as well as the interfering talker. The closer the value of $\rho_l[m]$ is to 1, the lower the effect of reverberation and interfering talkers on that portion of the signal. Thus the value of $\rho_l[m]$ is used in order to separate the portions of the signals that are severely affected by the reverberant field and the interfering noise by the use of a weight matrix.

Based on $\rho_l[m]$, the weight computation was given by,

$$w_l[m] = \rho_l[m]^a \tag{4.7}$$

The nonlinearity a is introduced to cause a sharp decay of w_l as a function of ρ_l and it was experimentally set to 3. The weights computed are applied as given below:

$$Y_l[n;m] = w_l[m] \bar{x}[n;m] \tag{4.8}$$

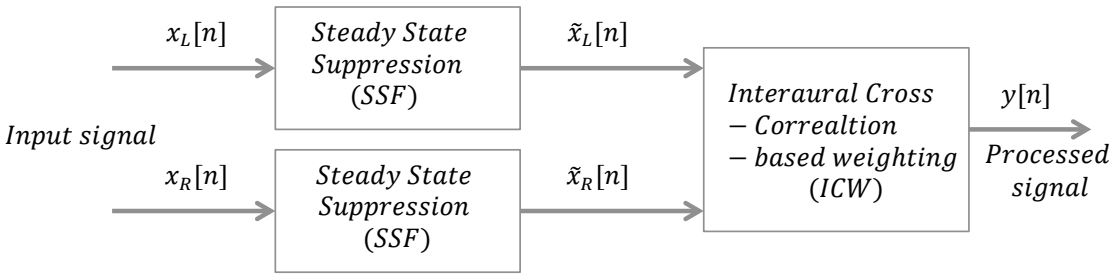


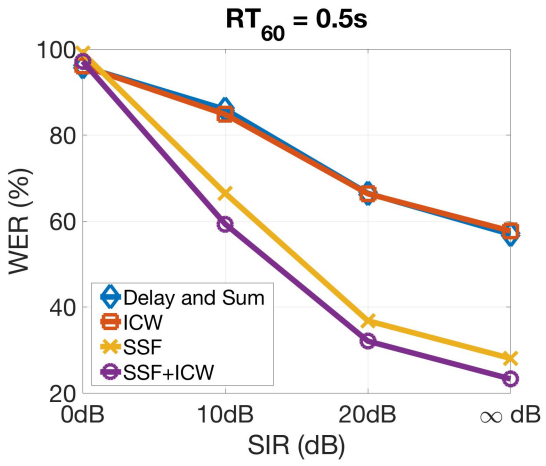
Figure 4.4: Overall block diagram of processing using steady-state suppression and interaural cross-correlation based weighting.

where $Y_l[n;m]$ is the short-time signal corresponding to the m^{th} frame and l^{th} gammatone channel and $\bar{x}[n;m]$ is the average of short-time signals $x_{R,l}[n;m]$ and $x_{L,l}[n;m]$ corresponding to the m^{th} frame and l^{th} gammatone channel. To resynthesize speech, all l channels are then combined.

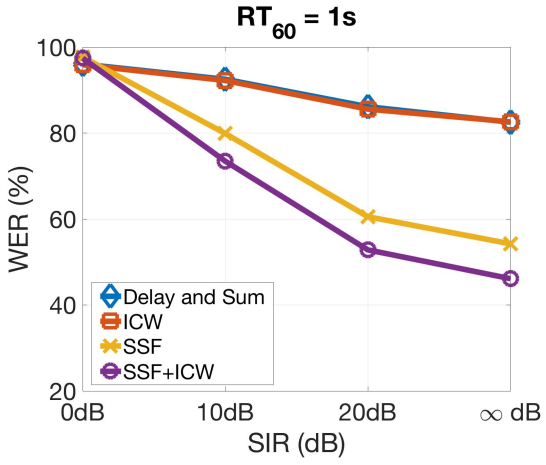
4.3 Experimental Results

The ICW algorithm was used in conjunction with the SSF algorithm in this study. The SSF algorithm leads to significant improvements in the presence of reverberation which in turn, leads to better ITD-based weighting. The overall block diagram is shown in Figure 4.4. Steady-state suppression, described in Section 2.5.1, is performed monaurally, and subsequently a weight that is based on interaural cross-correlation is applied to the signal, as described in Section 4.2. All experiments were conducted using simulated data as described in Chapter 3.

Preliminary results were obtained using the CMU Sphinx speech recognition system. The RM1 database was used for the preliminary experiments and reverberation times of 0.5 s and 1 s were used for this purpose. An interfering talker was mixed in at 0 dB, 10 dB and 20 dB SIR. The absence of an interfering talker was also included as one of the con-

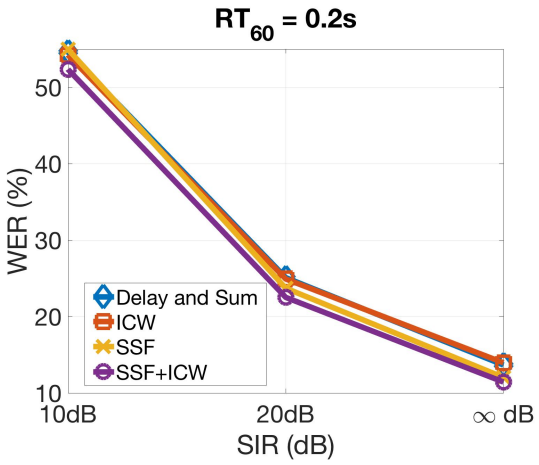


(a)

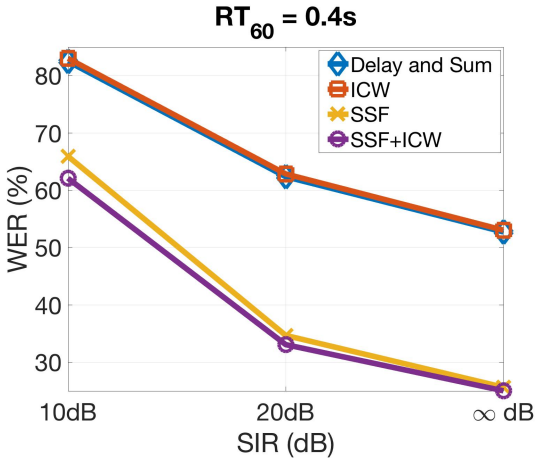


(b)

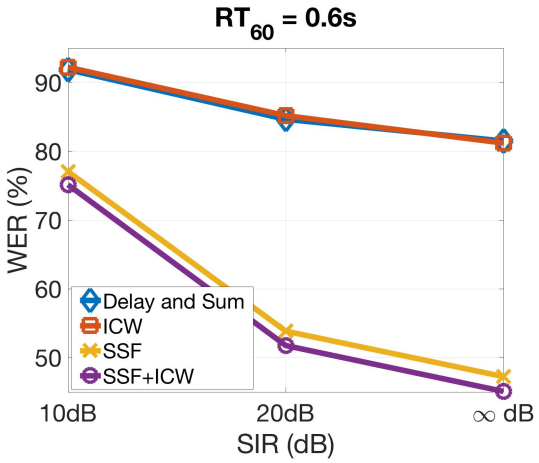
Figure 4.5: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the RM1 database and the CMU Sphinx speech recognition system using clean training data: (a) 0.5 s (b) 1 s.



(a)

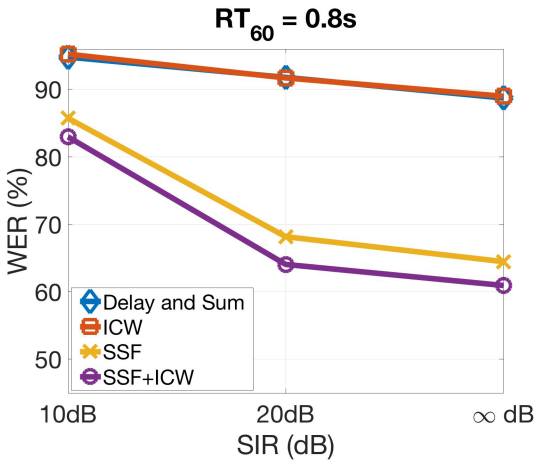


(b)

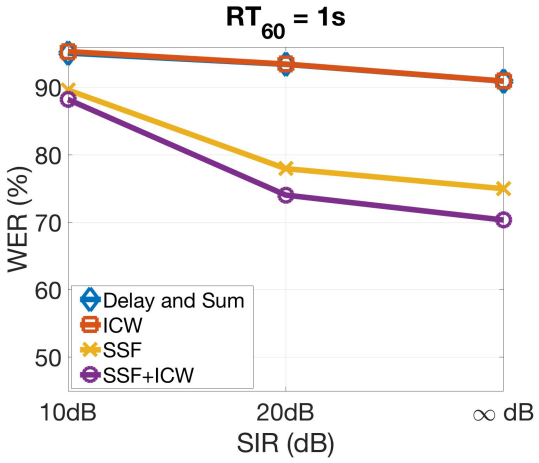


(c)

Figure 4.6: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a GMM-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.2 s (b) 0.4 s (c) 0.6s.



(a)



(b)

Figure 4.7: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a GMM-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.8 s (b) 1 s.

WER for $RT_{60} = 0.2s$	10 dB	20 dB	∞ dB
Delay and Sum	47.88%	21.07%	10.42%
ICW	48.23%	21.09%	10.63%
SSF	45.32%	16.08%	7.88%
SSF+ICW	41.98%	14.89%	7.81%

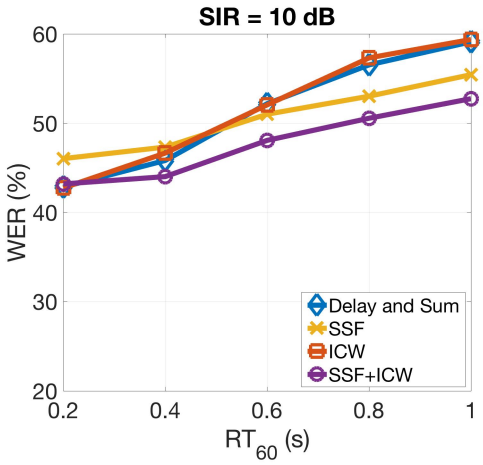
WER for $RT_{60} = 0.4s$	10 dB	20 dB	∞ dB
Delay and Sum	75%	51.49%	41.01%
ICW	75.68%	51.88%	41.71%
SSF	56.38%	26.1%	18.94%
SSF+ICW	53.8%	25.48%	17.78%

WER for $RT_{60} = 0.6s$	10 dB	20 dB	∞ dB
Delay and Sum	87.56%	73.12%	65.37%
ICW	88.03%	74.41%	67.07%
SSF	68.17%	42.54%	35.49%
SSF+ICW	66.6%	41.83%	34.9%

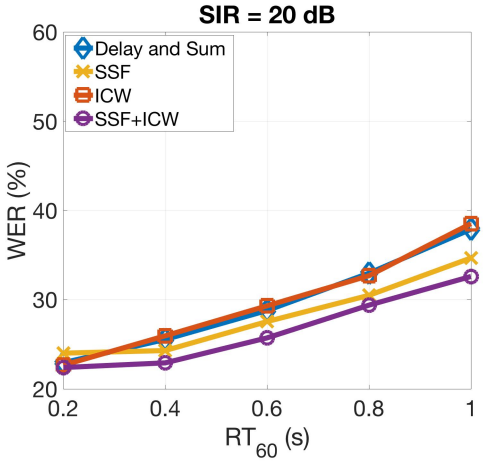
WER for $RT_{60} = 0.8s$	10 dB	20 dB	∞ dB
Delay and Sum	90.86%	82.63%	76.82%
ICW	90.83%	83.62%	77.88%
SSF	76.42%	54.98%	48.42%
SSF+ICW	75.96%	54.01%	49.04%

WER for $RT_{60} = 1s$	10 dB	20 dB	∞ dB
Delay and Sum	93.27%	86.34%	83.62%
ICW	93.95%	87.17%	83.64%
SSF	83.41%	65.44%	60.58%
SSF+ICW	82.29%	62.94%	57.82%

Table 4.1: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times of 0.2 s, 0.4 s, 0.6 s, 0.8 s and 1 s using the WSJ database and a DNN-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data.



(a)



(b)

Figure 4.8: Word Error Rate as a function of reverberation time at various Signal-to-Interference Ratios using the WSJ database and a DNN-based acoustic model obtained using multi-style training using the Kaldi speech recognition toolkit: (a) 10dB (b) 20dB.

ditions. More detailed experiments were later conducted for reverberation times of 0.2 s to 1 s in steps of 0.2 s using the Kaldi speech recognition toolkit and the WSJ database. In this case, an interfering talker was mixed in at signal-to-interference ratios (SIRs) of 10 dB and 20 dB. Experiments were also conducted in the absence of an interfering talker. These experiments were conducted using both GMM-based and DNN-based acoustic models trained using Kaldi. Clean as well as multi-style training were used. In cases where SSF was used in the test algorithm, the training data also was processed using the SSF algorithm.

For multi-style training, the training data contained roughly equal number of utterances with simulated reverberation times of 0.25, 0.5 and 0.75 s. The location of the microphone setup was randomized for each utterance. The training data utterances did not contain any interfering talkers. Results were reported using the test data at 10 and 20 dB SIR for reverberation times of 0.2 s to 1 s in steps of 0.2 s.

Figures 4.5, 4.6, 4.7, 4.8 and Table 4.1 show the results obtained using baseline Delay-and-Sum processing, the SSF algorithm alone, the ICW algorithm alone, and the combination of the SSF and ICW algorithms. The performance of the SSF+ICW algorithm is compared to that of SSF alone (monaurally) and ICW alone. The results of the Delay-and-Sum algorithm serve as baseline.

As seen in Figure 4.5, the ICW algorithm applied by itself does not provide any improvement in performance compared to baseline Delay-and-Sum processing. Nevertheless, the addition of ICW to SSF does lead to a reduction in WER compared to performance obtained using SSF alone as seen in Figure 4.5. While the WER remains the same for 0 dB SIR, for all the other conditions, the addition of ICW to SSF decreases the relative WER by up to 17%. There is a consistent improvement in WER for 10 dB and 20 dB SIR and in the absence of an interfering talker. The inclusion of envelope ITD cues and their coherence across binaural signals helps reduce the effects of both interfering noise and reverberation.

These trends remain consistent even with the use of the Kaldi speech recognition sys-

tem as seen in Figures 4.6, 4.7 and Table 4.1. While the absolute numbers vary quite a bit, as expected, an improvement is seen in WER. In case of the GMM-based acoustic model using the WSJ database, the relative improvements in WER remain more or less constant across SIR in the case of lower reverberation times as seen in Figure 4.6a. As reverberation increases, the improvements in the presence of higher noise diminish. The WER in these cases is quite high to begin with even after the application of SSF which would explain the lower improvements. The results for the DNN-based acoustic model are similar. Table 4.1 shows results for the Kaldi DNN-based acoustic model for reverberation times of 0.2 s to 1 s. The lowest WER for each SIR condition is highlighted. Clean training was used for these results.

The combination of SSF and ICW algorithms leads to an improvement in WER compared to using SSF alone as seen in Table 4.1. Lower improvements are seen in cases where the WER is already quite low or in noisier conditions when the WER is very high.

Similar trends are seen for the results obtained using multi-style training in Figure 4.8. As expected, the WER gets significantly better with the use of multi-style training for all the algorithms. However, the use of ICW in conjunction with SSF still gives the best performance across different reverberation times. In fact, the relative improvement seen by the addition of ICW to SSF compared to using SSF alone remains fairly consistent across reverberation times and SIR.

4.4 Conclusions

The ICW algorithm weighs the contributions of different frames of speech according to the extent to which the amplitude envelopes in sub-band frequencies are correlated across two microphones, which should serve as a measure of the extent to which the signals are locally unaffected by the effects of reverberation. The use of the ICW algorithm in combi-

nation with the SSF algorithm leads to improvements in WER in the presence of reverberation and interfering noise. Using signal envelopes for ITD-based processing does lead to better speech recognition. Relative improvements with ICW seem best in the presence of moderate to high reverberation and moderate interfering noise while using clean traning data. In the case of multi-style training, the improvements with ICW remain more or less similar across different reverberation times and interfering noise levels.

CHAPTER 5

BINAURAL PROCESSING USING INTER-MICROPHONE COHERENCE

The goal of coherence-based processing is to place greater emphasis on signal components that appear to be coherent across microphones. The ICW algorithm discussed in Chapter 4 uses across-microphone ITD-based processing to compute a weight matrix. This goal of this weight matrix is to separate portions of the signal that are not coherent and thus presumably, do not originate from the target source. No assumptions are made about the nature of the microphone signals or the noise that may be present in the room. In this section, a different method of coherence-based processing is discussed that is based on a model of reverberated and noisy speech.

In the presence of multiple microphones, where spatial information is available, one proposed approach to mitigate the effects of reverberation and noise is to be able to characterize each portion of the speech signal as dominated by coherent or diffuse energy. A technique proposed in [31] uses spectral subtraction initially for suppression of late reverberations followed by coherence-based processing.

Other model-based approaches may also be used to determine the degree of coherence between the two microphone signals [32, 33]. In such a case, models of coherence for

a coherent sound field versus a diffuse sound field are developed. For a given signal, it is useful to understand how coherent or diffuse different portions of the signal are. With this knowledge it is possible to apply a mask to the input signal that suppresses regions where the ratio of coherent-to-diffuse energy is low. The Coherent-to-Diffuse Ratio-based Weighting (CDRW) described in Section 5.1, uses this principle to reject regions of the speech signal that are not coherent. This technique, in conjunction with steady state suppression, is used to improve ASR in the presence of reverberation and interfering talkers.

5.1 Interaural coherence-based processing

One of the earliest approaches using interaural coherence was proposed by Allen *et al.* [34] where different gain factors were applied to different parts of the signal to suppress components that were mainly reverberant. The computation of the gain factors was performed by determining the diffuseness of the sound field between the microphones. Several variations of this technique were proposed including binaural application of the gain factors as well as the inclusion of head related transfer functions [35]. Westermann *et al.* [36] also extended the concepts introduced in [34] to make use of interaural coherence histograms for binaural dereverberation.

There are several possible approaches for computation of gain or weights that best eliminate reverberation and noise. In this work, we make use of the Coherent-to-Diffuse Ratio (CDR) for weight computation. As the name suggests, this metric provides an estimate of the extent to which a particular portion of the signal is affected by the reverberation or noise present. We use the method proposed by Jeub *et al.* [32] for deriving the Coherent-to-Diffuse Ratio (CDR).

5.2 Structure of the CDRW algorithm

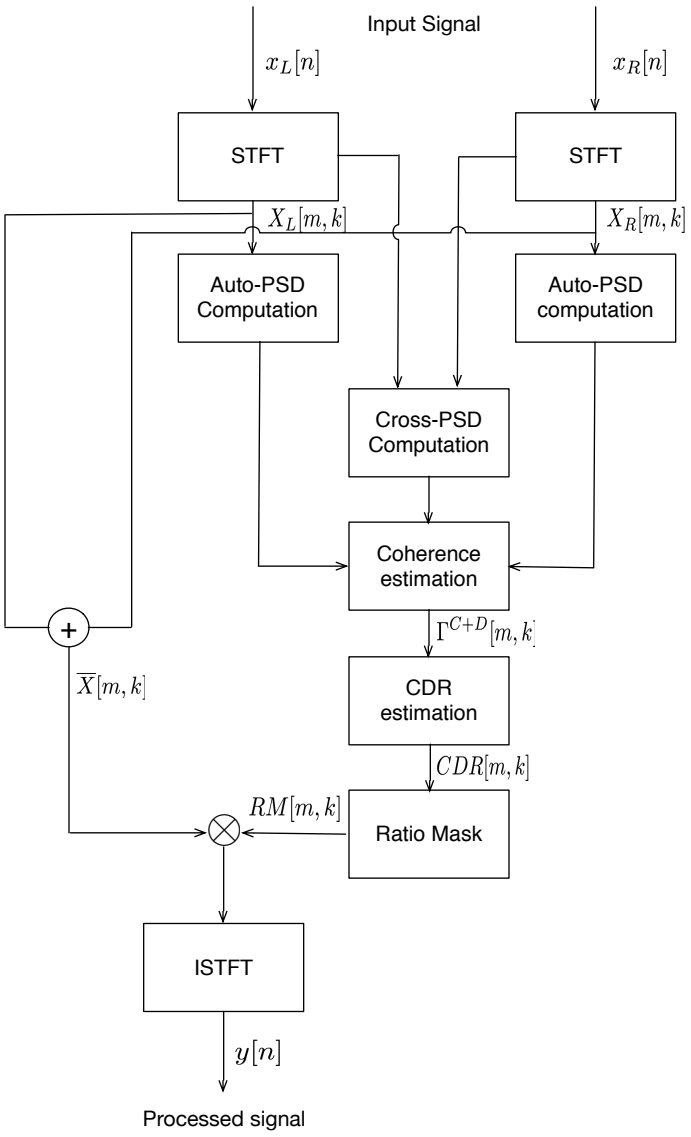


Figure 5.1: Block diagram describing the CDRW algorithm.

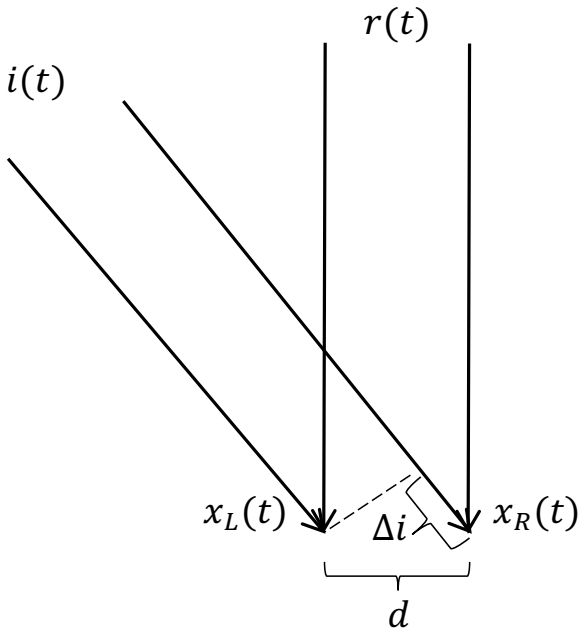


Figure 5.2: Diagram depicting the microphone setup used. Signals $x_R(t)$ and $x_L(t)$ are the right and left microphones respectively that capture sounds coming from the target $r(t)$ and the interferer $i(t)$ in a reverberant room.

A block diagram describing the Coherent-to-Diffuse Ratio-based Weighting algorithm (CDRW) used in this work is shown in Figure 5.1.

5.2.1 Coherence function

Consider two signals from the microphones $x_R(t)$ and $x_L(t)$ as seen in Figure 5.2. The coherence function is a statistic commonly used to measure how correlated two signals are. The complex interaural coherence $\Gamma_{x_R x_L}(\omega)$ between the signals $x_R(t)$ and $x_L(t)$ is given by

$$\Gamma_{x_R x_L}(\omega) = \frac{\Phi_{x_R x_L}(\omega)}{\sqrt{\Phi_{x_R x_R}(\omega)\Phi_{x_L x_L}(\omega)}} \tag{5.1}$$

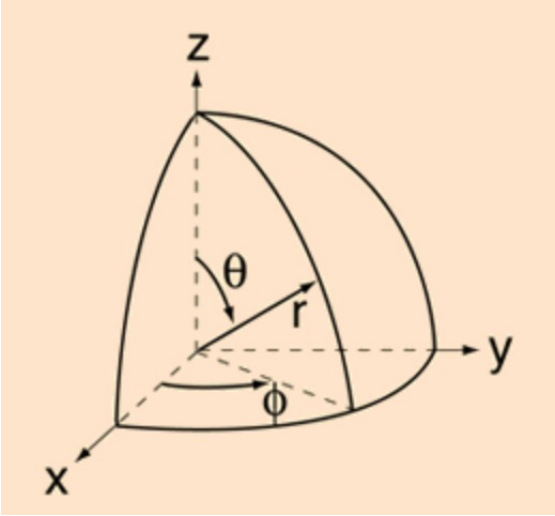


Figure 5.3: Portion of a sphere with radius r.

where $\Phi_{x_R x_L}(\omega)$ denotes the Cross-Power Spectral Density (Cross-PSD) of $x_R(t)$ and $x_L(t)$ and $\Phi_{x_R x_R}(\omega)$ and $\Phi_{x_L x_L}(\omega)$ denote the Auto-PSDs of $x_R(t)$ and $x_L(t)$ respectively. The Auto-PSD and Cross-PSD functions can be estimated using recursive averaging.

5.2.2 Coherence in a diffuse field

In the case of a diffuse field, as is caused by reverberation, the spherically-isotropic interaural coherence can be calculated by integrating all the plane waves originating from a surface area over the whole surface area of a sphere [37]. For a spherical surface as shown in Figure 5.3, it is assumed that the microphones are placed on the x-axis. It is also assumed that the distance d between the microphones is much smaller than the radius r. Azimuth ϕ and elevation θ are also shown in Figure 5.3. On the given surface of the sphere, an infinitesimal area is given by $dA = r^2 sin\phi d\phi d\theta$.

The coherence between two signals can be computed by integrating over all plane waves originating from some small surface area A.

$$\Gamma^D(\omega) = \frac{\oint_A \Phi_{x_R x_L}(\omega) dA}{\oint_A \sqrt{\Phi_{x_R x_R}(\omega)\Phi_{x_L x_L}(\omega)} dA} \tag{5.2}$$

Consider first the case of a single plane wave originating from some angle ϕ. A simplified 2-D illustration is shown in Figure 5.2.

As seen in Equation 4.1, this leads to a delay between the signal reaching the two sensors. This delay τ owing to path difference is

$$\tau = \frac{dcos\phi}{c} \tag{5.3}$$

where d is the distance between the two sensors and c is the speed of sound. Thus the cross-power spectral density is,

$$\Phi_{x_R x_L}(\omega) = \Phi_{x_R x_R}(\omega) e^{\frac{-j\omega dcos\phi}{c}} \tag{5.4}$$

Given the isotropic assumption, the power spectral densities are independent of location.

$$\Phi_{x_L x_L}(\omega) = \Phi_{x_R x_R}(\omega) \tag{5.5}$$

The spatial coherence of the isotropic diffuse field can now be calculated by integrating over all plane waves as seen in Equation 5.2.

$$\begin{aligned} \Gamma^D(\omega) &= \frac{\oint_A \Phi_{x_R x_R}(\omega) e^{\frac{-j\omega dcos\phi}{c}} dA}{\oint_A \Phi_{x_R x_R}(\omega) dA} \\ &= \frac{1}{A} \oint_A e^{\frac{j\omega dcos\phi}{c}} dA \end{aligned}$$

The area A on the surface of the sphere is $A = 4\pi r^2$ and the integral can be computed over $\phi \in [0, \pi]$ and $\theta \in [0, 2\pi)$.

$$\begin{aligned}\Gamma^D(\omega) &= \frac{1}{4\pi r^2}\int_0^{2\pi}\int_0^{\pi} e^{\frac{-j\omega dcos\phi}{c}} r^2 sin\phi d\phi d\theta \\ &= \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} e^{\frac{-j\omega dcos\phi}{c}} sin\phi d\phi d\theta\end{aligned} \tag{5.6}$$

Substituting $g = \frac{\omega dcos\phi}{c}$,

$$\begin{aligned}\Gamma^D(\omega) &= \frac{c}{2\omega d}\int_{\frac{-\omega d}{c}}^{\frac{\omega d}{c}} e^{-jg} dg \\ &= \frac{sin(\omega d/c)}{\omega d/c} \\ &= sinc(\frac{\omega d}{c})\end{aligned} \tag{5.7}$$

In the case of a coherent source with the signal arriving at some angle ϕ, the two signals only differ by a delay term and the interaural coherence is given by,

$$\Gamma^C_{x_R x_L}(\omega) = e^{-\left(\frac{j\omega dcos\phi}{c}\right)} \tag{5.8}$$

5.2.3 Coherent-to-Diffuse Ratio

The Coherent-to-Diffuse Ratio (CDR) is defined as the ratio of the coherent energy to the diffuse energy in a given environment. The expression for CDR in Equation 5.16 was derived by Jeub and his colleagues [32].

$$CDR(\omega) = \frac{\Phi^C(\omega)}{\Phi^D(\omega)} \tag{5.9}$$

where $\Phi^C(\omega)$ and $\Phi^D(\omega)$ denote the auto-PSD corresponding to a coherent and diffuse source respectively.

In the case of a diffuse field, the noise signals at both microphones are assumed to have equal power spectral density denoted by $\Phi^D(\omega)$. Using the relationship as seen in Equation 5.1, the cross-PSD can be derived as,

$$\Phi_{x_L x_R}(\omega) = \Phi^D(\omega) sinc(\frac{\omega d}{c})$$

Similarly, for a coherent source, the auto-PSD is $\Phi^C(\omega)$. Again, using Equation 5.1, the cross-PSD can be derived as,

$$\Phi_{x_L x_R}(\omega) = \Phi^C(\omega) e^{-\left(\frac{j\omega d cos\phi}{c}\right)}$$

Thus, in an environment involving a mix of diffuse and coherent sources, the interaural coherence can be given as a ratio of the the total cross-PSD to the total auto-PSD of the coherent and diffuse sounds.

$$\Gamma^{C+D}_{x_L x_R}(\omega) = \frac{\Phi^D(\omega) sinc(\frac{\omega d}{c}) + \Phi^C(\omega) e^{-\left(\frac{j\omega d cos\phi}{c}\right)}}{\Phi^C(\omega) + \Phi^D(\omega)} \tag{5.10}$$

Assuming the coherent source is located such that there is no delay between the two microphones i.e. $\phi = \pi/2$,

$$\Gamma^{C+D}_{x_L x_R}(\omega) = \frac{\Phi^D(\omega) sinc(\frac{\omega d}{c}) + \Phi^C(\omega)}{\Phi^C(\omega) + \Phi^D(\omega)} \tag{5.11}$$

Substituting Equation 5.9 into Equation 5.11,

$$\Gamma^{C+D}_{x_L x_R}(\omega) = \frac{sinc(\frac{\omega d}{c}) + CDR(\omega)}{CDR(\omega) + 1} \tag{5.12}$$

An expression for the CDR can be obtained by rearranging the terms in Equation 5.12.

$$CDR(\omega) = \frac{sinc(\frac{\omega d}{c}) - \Gamma^{C+D}_{x_L x_R}(\omega)}{\Gamma^{C+D}_{x_L x_R}(\omega) - 1} \tag{5.13}$$

The real-valued CDR can be given by,

$$CDR(\omega) = \frac{sinc(\frac{\omega d}{c}) - Re\{\Gamma^{C+D}_{x_L x_R}(\omega)\}}{Re\{\Gamma^{C+D}_{x_L x_R}(\omega)\} - 1} \tag{5.14}$$

To make sure that value of the CDR remains greater than 0,

$$CDR(\omega) = max\left(0, \frac{sinc\left(\frac{\omega d}{c}\right) - Re\{\Gamma^{(C+D)}_{x_R x_L}(\omega)\}}{Re\{\Gamma^{(C+D)}_{x_R x_L}(\omega)\} - 1}\right) \tag{5.15}$$

where d is the distance between the two microphones, c is the speed of sound and $\Gamma^{(C+D)}_{x_R x_L}(\omega)$ is the interaural coherence for a mixed (coherent+diffuse) source.

For the m^{th} frame and k^{th} frequency index, this can be expressed as,

$$CDR[m,k] = max\left(0, \frac{sinc\left(\frac{2\pi k f_s d}{Nc}\right) - Re\{\Gamma^{C+D}_{x_R x_L}[m,k]\}}{Re\{\Gamma^{C+D}_{x_R x_L}[m,k]\} - 1}\right) \tag{5.16}$$

The quantity of $\Gamma^{C+D}_{x_R x_L}[m,k]$ is estimated using recursive smoothing.

5.2.4 Mask estimation

Equation 5.16, as derived in [32], is useful in separating portions of the signal STFT that are dominated by the diffuse noise and therefore need to be suppressed. In order to convert the CDR quantity into a ratio mask, we use the classical Wiener filter.
For a signal $x[n]$ corrupted by noise $s[n]$, the resulting noisy signal is given by $y[n]$.

$$y[n] = x[n] + s[n] \tag{5.17}$$

Wiener filtering produces a minimum mean-squared error estimate of the clean signal $x[n]$ from $y[n]$. For the output of the Wiener filter denoted by of $\hat{x}[n]$, the error $e[k]$ is expressed as,

$$e[k] = \hat{x}[n] - x[n]$$

The Wiener filter thus aims to minimize the mean square error $E[|e[k]|^2]$. The transfer function of a Wiener filter is given by,

$$H_w(\omega) = \frac{\phi_{xy}(\omega)}{\phi_{yy}(\omega)} \tag{5.18}$$

where $\phi_{xy}(\omega)$ is the cross power spectral density between the signal $x[n]$ and the noisy signal $y[n]$ and $\phi_{yy}(\omega)$ is the auto power spectral density of the noisy signal $y[n]$.

Considering Equation 5.17 and assuming that $x[n]$ and $s[n]$ are not correlated,

$$\phi_{yy}(\omega) = \phi_{xx}(\omega) + \phi_{ss}(\omega)$$

and

$$\phi_{xy}(\omega) = \phi_{xx}(\omega)$$

Substituting the expressions for $\phi_{xy}(\omega)$ and $\phi_{yy}(\omega)$ into Equation 5.18,

$$\begin{aligned} H_w(\omega) &= \frac{\phi_{xy}(\omega)}{\phi_{yy}(\omega)} \\ &= \frac{\phi_{xx}(\omega)}{\phi_{xx}(\omega) + \phi_{ss}(\omega)} \end{aligned} \tag{5.19}$$

We can define a frequency-dependent Signal-to-Noise-Ratio as $SNR(\omega) = \frac{\phi_{xx}(\omega)}{\phi_{ss}(\omega)}$.

$$H_w(\omega) = \frac{SNR(\omega)}{SNR(\omega) + 1} \tag{5.20}$$

Thus, the Wiener filter transfer function can be expressed in terms of $SNR(\omega)$. In place of $SNR(\omega)$, we use $CDR(\omega)$ in this study to produce a ratio mask $RM(\omega)$. CDR is an SNR-like measure in the sense that it provides a ratio of the coherent-to-diffuse power which

effectively is like the ratio of desired signal to noise power.

$$RM(\omega) = \frac{CDR(\omega)}{CDR(\omega) + 1} \tag{5.21}$$

For the m^{th} frame and k^{th} frequency index, this can be expressed as,

$$RM[m, k] = \frac{CDR[m, k]}{CDR[m, k] + 1} \tag{5.22}$$

The ratio mask is applied to the STFT of the mean of the two microphone inputs and an Inverse STFT (ISTFT) is then performed to obtain the processed waveform. In this study, the combination of CDRW and SSF gave the best performance in terms of ASR. An example of the spectrograms of the original waveform and the waveform after processing using the SSF+CDRW algorithm is shown in Figure 5.4.

5.3 Experimental Results

Experiments were conducted for reverberation times of 0.2 s to 1 s in steps of 0.2 s using the Kaldi speech recognition toolkit [38] and the WSJ database [39]. All experiments were conducted using simulated data. An interfering talker was mixed in at 10 dB and 20 dB. Experiments were also conducted in the absence of an interfering talker. These experiments were conducted using both GMM-based and a DNN-based acoustical model trained using Kaldi. The DNN-based model used alignments generated using the GMM-based models. The DNN-based model has two hidden layers. Clean as well as multi-style training was performed. Training data underwent processing identical to the test data.

In the case of multi-style training, the training data contained roughly equal number of utterances with simulated reverberation times of 0.25, 0.5 and 0.75 s. The location of the microphone setup was randomized for each utterance. The training data utterances did

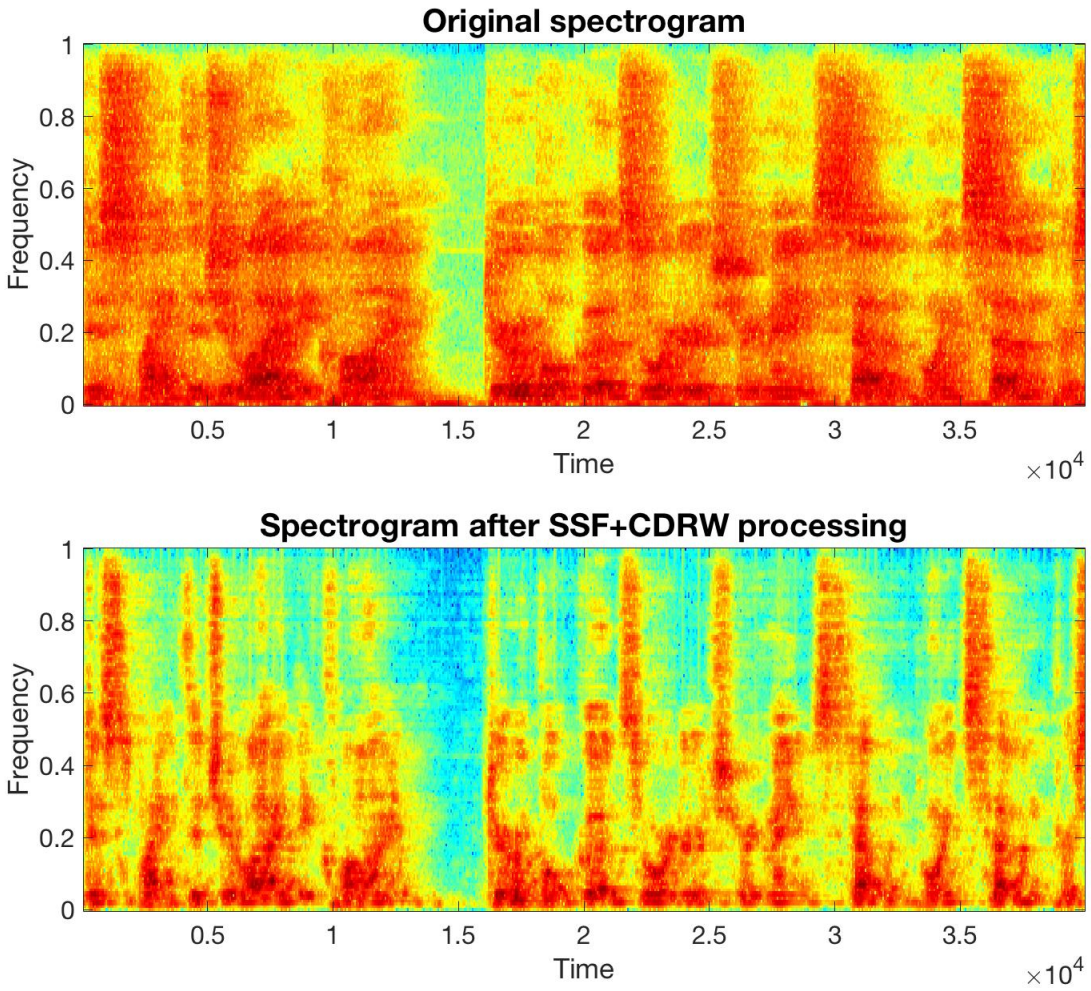


Figure 5.4: SSF+CDRW processing on WSJ utterance at reverberation time of $RT_{60} = 0.6s$ in the absence of any interferer (a) Original spectrogram (b) Spectrogram after SSF+CDRW processing.

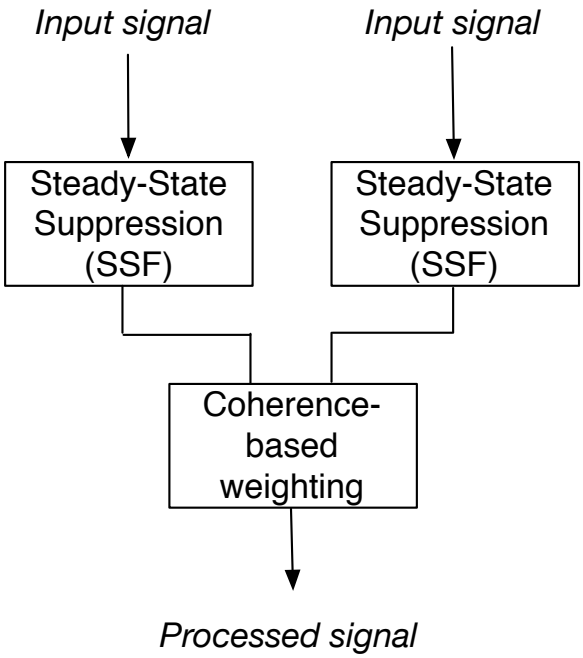
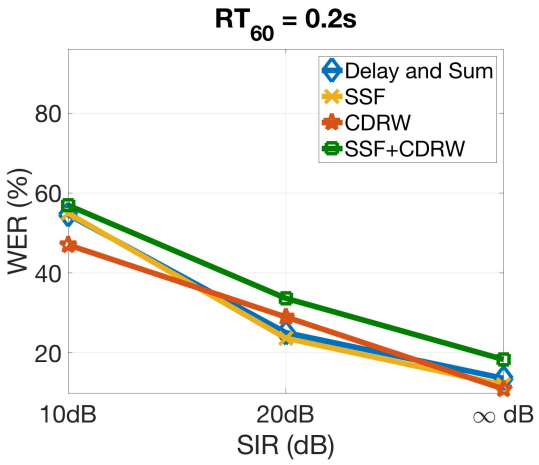


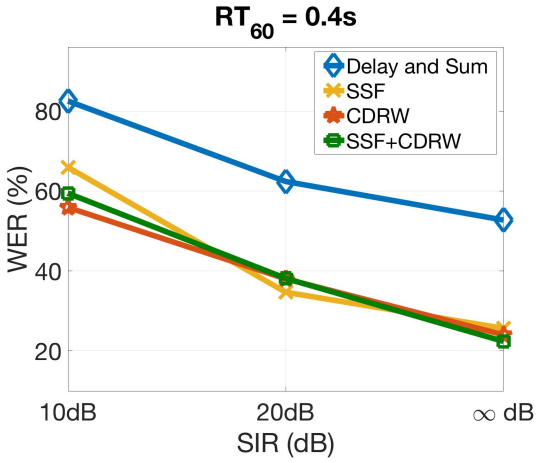
Figure 5.5: A block diagram of SSF+CDRW processing. SSF is performed monaurally on the signals from the right and left sensor after which CDRW is applied.

not contain any interfering talkers. Results were reported using the test data at 10 and 20 dB SIR for reverberation times of 0.2 s to 1 s in steps of 0.2 s.

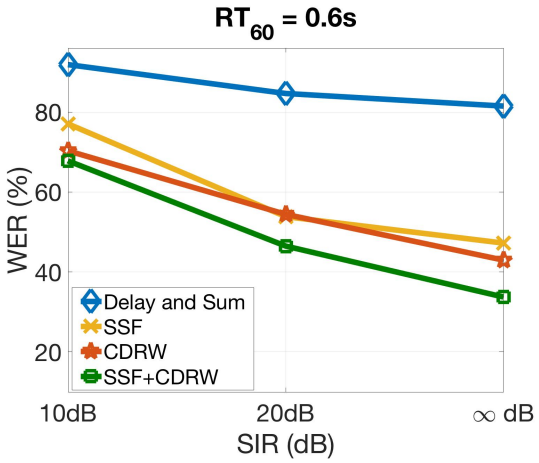
Results obtained using steady-state suppression alone (monaurally) are compared to results using the method introduced in this chapter which includes the CDRW algorithm and SSF processing followed by CDRW. A block diagram for how SSF+CDRW processing was performed is shown in Figure 5.5. As seen, the SSF algorithm is performed monaurally for the right and left microphone. The CDRW algorithm is then applied to the output of the SSF algorithm. This is done for reverberant and noisy environments. In the case of low reverberation times as seen in Figure 5.6a, using the Kaldi GMM-based model, the CDRW algorithm alone performs pretty well and has the lowest WER at 10 dB SIR and in the absence of an interferer. The SSF+CDRW method does not help in this case. In fact, the combination of SSF+CDRW leads to WER that is worse than using just Delay and Sum at reverberation time of 0.2 s. As the reverberation time increases, the improvements due



(a)

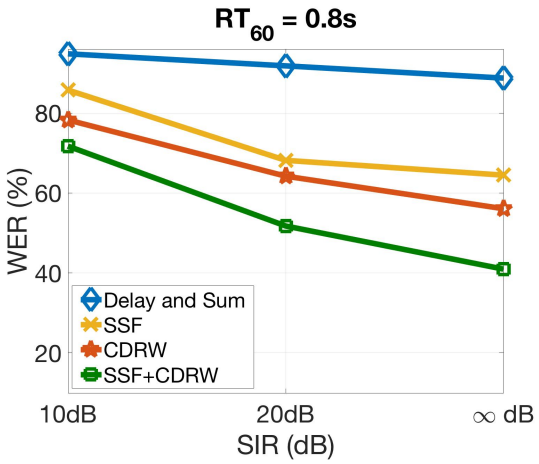


(b)

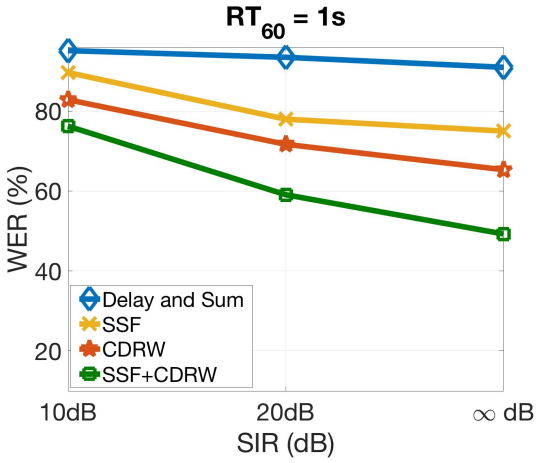


(c)

Figure 5.6: Word Error Rate as a function of the Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a GMM-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.2 s (b) 0.4 s (c) 0.6s.

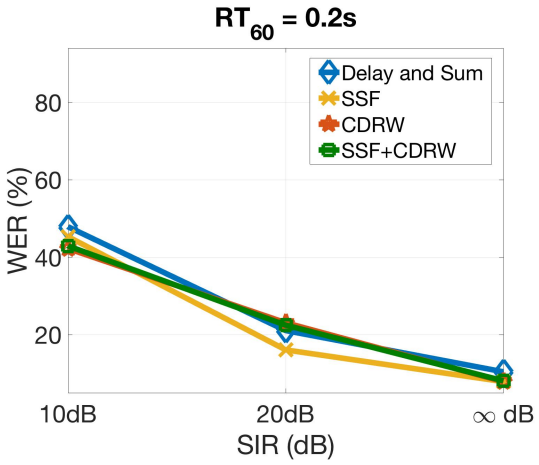


(a)

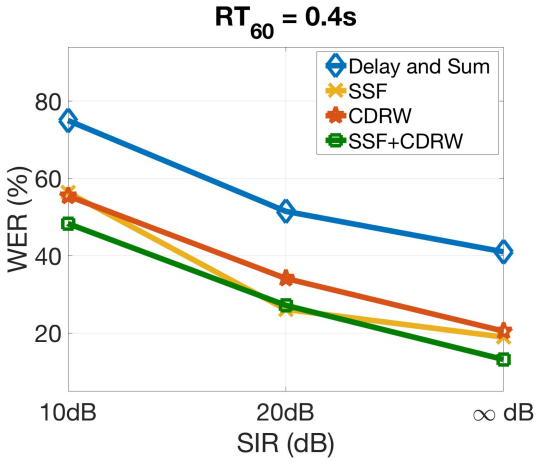


(b)

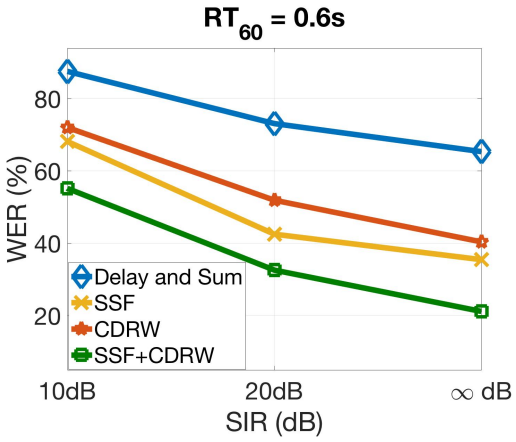
Figure 5.7: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a GMM-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.8 s (b) 1 s.



(a)

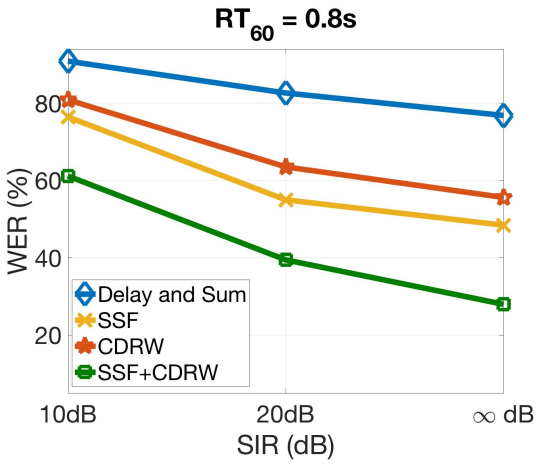


(b)

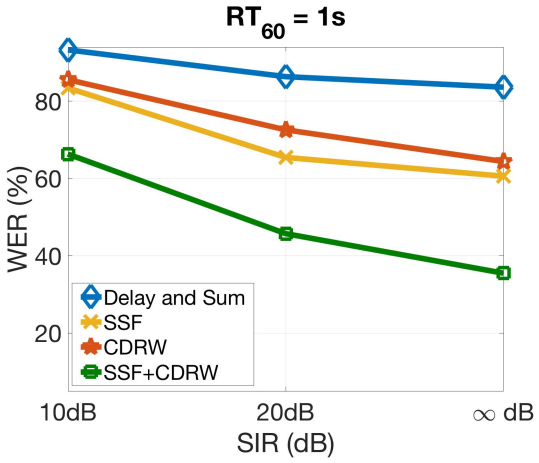


(c)

Figure 5.8: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a DNN-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.2 s (b) 0.4 s (c) 0.6s.

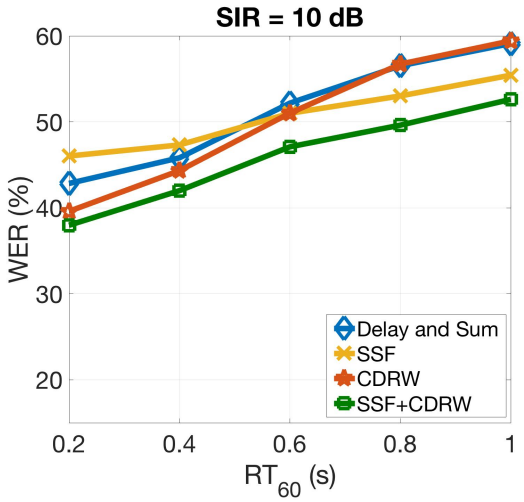


(a)

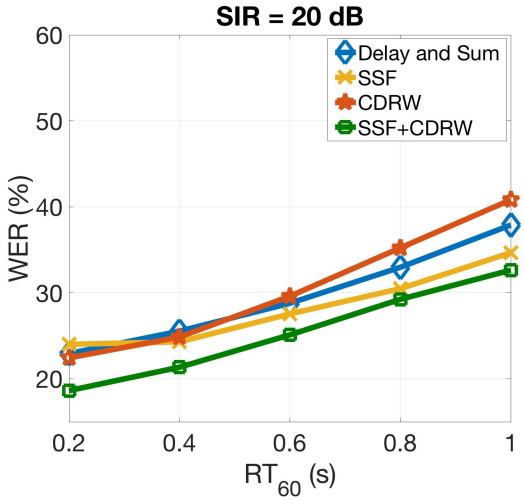


(b)

Figure 5.9: Word Error Rate as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at various reverberation times using the WSJ database and a DNN-based acoustic model trained using the Kaldi speech recognition toolkit using clean training data: (a) 0.8 s (b) 1 s.



(a)



(b)

Figure 5.10: Word Error Rate as a function of reverberation time at various SIRs for the WSJ database and a DNN-based acoustic model obtained using multi-style training using the Kaldi speech recognition toolkit: (a) 10dB (b) 20dB.

to SSF+CDRW are seen. The CDRW algorithm leads to lower WER compared to SSF at reverberation times of 0.4 s, 0.6 s, 0.8 s and 1 s and at these conditions the combination of SSF+CDRW algorithms leads to the lowest WER. At reverberation time of 0.6 s, there is a relative improvement in WER of nearly 29% at ∞ dB which is very significant. These trends continue as the reverberation gets worse.

The results using the DNN-based models trained using the Kaldi speech recognition toolkit and clean training data are seen in Figures 5.8 and 5.9. While the CDRW algorithm leads to better performance compared to SSF as seen in Figures 5.6 and 5.7 using the GMM-based model, this is no longer the case while using the DNN-based model as seen in Figures 5.8 and 5.9. At lower reverberation times, the improvements in WER using CDRW are limited. For higher reverberation times, even though the CDRW algorithm does much better than the baseline Delay and Sum system, it still does not do better than the SSF algorithm. The combination of SSF and CDRW algorithms gives the best performance for reverberation times of greater than 0.2 s. In fact, for a reverberation time of 0.6 s, a relative improvement in WER of over 40% is seen at ∞ dB. This stays consistent at reverberation times of 0.8 and 1 s as well.

Overall, the combination of SSF+CDRW algorithm does not lead to an improvement in WER at low reverberation times close to 0.2 s. However, as the reverberation increases, the improvements increase significantly and stay consistent even for reverberation time of 1 s.

The results using multi-style training with the Kaldi speech recognition toolkit are seen in Figure 5.10. At 10 dB SIR, the CDRW algorithm leads to lower error than the SSF algorithm for low to moderate reverberation. However, at reverbertiom times of 0.6 s or higher, the CDRW algorithm leads to much higher WER compared to SSF. The combination of SSF with CDRW leads to the lowest WER across the board. This is true for both the interfering noise levels that were tested. The relative improvement in WER seen by the addition of CDRW to SSF is greater at lower reverberation times, however.

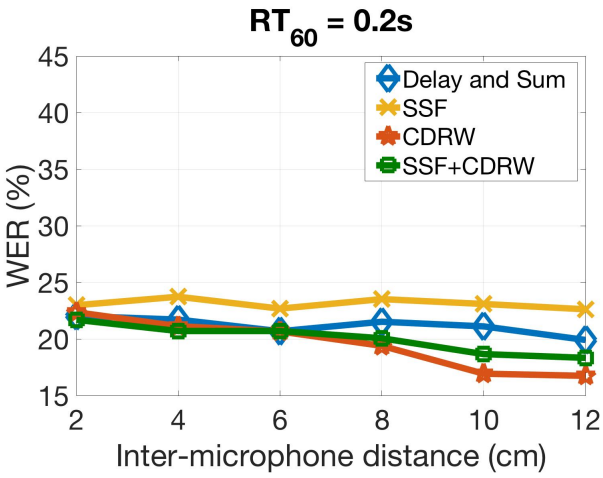
5.4 Effect of inter-microphone distance

The CDR metric depends, among other factors, on the inter-microphone distance. In particular, as seen in Equation 5.7, inter-microphone distance plays a role in the determining the coherence in the spatially isotropic case, which in turn leads to the computation of the CDR metric. In order to study the effect of changing inter-microphone distance on the results, further experiments were conducted.

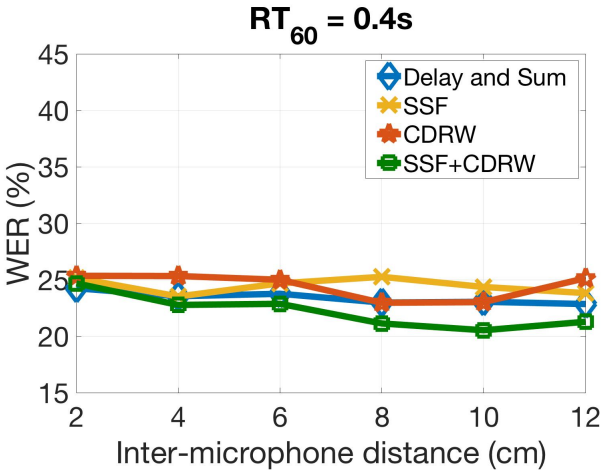
For these experiments, a new training set was generated that not only had utterances at reverberation times of 0.25, 0.5 and 0.75 s, but also had roughly equal utterances simulated at inter-microphone distances of 2.5, 5 and 7.5 cms. Test data had utterances simulated with inter-microphone distances of 2, 4, 6, 8, 10 and 12 cms. Test data with an interfering talker mixed in at 20 dB SIR was used for these experiments. A DNN-based model was trained using the Kaldi speech recognition toolkit. The results obtained are seen in Figure 5.11.

At the reverberation time of 0.2 s, as seen in Figure 5.11, the WER using the CDRW algorithm and by extension the SSF+CDRW algorithm drops with the increase in inter-microphone distance. Since the inter-microphone distance indirectly determines the range of frequencies over which the CDR metric is effective, this is not surprising. As seen in Equation 5.7, the frequency up to which the diffuse field is highly coherent is inversely proportional to the inter-microphone distance d. This would mean that at higher inter-microphone distances, the CDRW algorithm is effective for the most part of the signal.

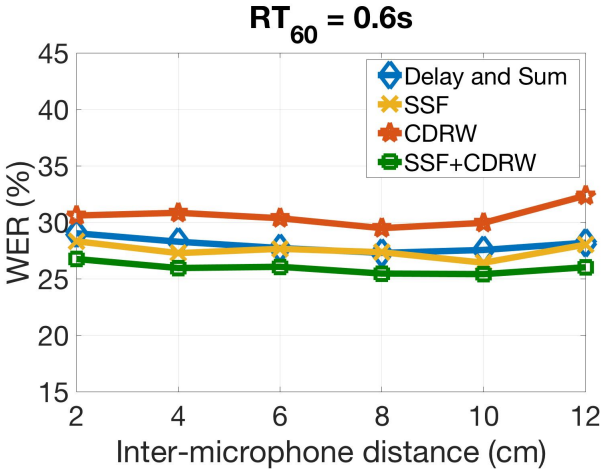
However, for the greater reverberation times of 0.6 s and over, the WER using the CDRW algorithm no longer decreases with increasing inter-microphone distance. The WER using the SSF+CDRW algorithm remains more or less the same across inter-microphone distance for reverberation times of 0.6 s and above. This could be because spatial aliasing leads to greater (worse) WER which becomes more prominent as the reverberation in-



(a)



(b)



(c)

Figure 5.11: Word Error Rate as a function of inter-microphone distance at various reverberation times. The WSJ database and a DNN-based acoustic model obtained using multi-style training using reverberated speech simulated using different inter-microphone distances was used: (a) 0.2s (b) 0.4s (c) 0.6s.

creases. Results for reverberation times greater than 0.6 s are not shown but follow trends that are very similar to Figure 5.11c.

5.5 Conclusions

In this chapter we introduce a novel method that provides better speech recognition accuracy in reverberant and noisy conditions. The approach uses CDR to derive a weight matrix that suppresses portions of the signal dominated by noise and reverberation. This method by itself and in combination with the SSF algorithm has been tested in this chapter. The combination of the SSF and CDRW algorithms leads to improvements of up to 42% relative in WER using the DNN-based acoustical models in Kaldi obtained using clean training. It is to be noted that neither the SSF nor the CDRW algorithms actively suppress the interfering signal, which is why the relative improvements obtained using the SSF and CDRW algorithms individually and in combination is greater in the absence of an interferer across different reverberation times.

CHAPTER 6

COHERENCE ACROSS FREQUENCY

Chapters 4 and 5 introduce methods to mitigate the effect of reverberation and interfering noise using coherence across the microphone signals. In addition to looking at coherence across the signals at the two sensors, it is also possible to leverage coherence seen in other domains within the same signals. Signals that arrive at the two microphones at the same time normally exhibit coherence in arrival time over a range of frequencies. It can be beneficial to capture this coherence in order to isolate signals coming from a source of interest.

One way to do this is to perform cross-correlation over some range of frequencies, which is the approach we adopt in this work. One of the earliest models of binaural hearing was proposed by Sayers and Cherry [40], which related the lateralization of binaural signals to their interaural cross-correlation. In binaural speech processing, a popular approach towards isolating target sounds in adverse environments is the grouping of sources according to common source location. This usually entails the use of interaural time difference (ITD) and interaural intensity difference (IID) as discussed in Chapters 4 and 5. Models that describe how these cues are used to lateralize sound sources are reviewed in [41, 42], among other sources.

Straightness weighting refers to a hypothesis that greater emphasis in auditory lateralization is given to the contributions of ITDs that are consistent over a range of frequencies

[43, 44, 45]. This was motivated by the fact that sounds emitted by point sources produced ITDs that are consistent over a range of frequencies. Hence, the existence of a "straight" maximum of the interaural cross-correlation function over a range of frequencies could be used to identify the correct ITD. In this chapter, we introduce a new method based on this concept called the Cross-Correlation across Frequency algorithm (CCF) . In essence, this method aims at boosting regions of coherence across frequency, and it also provides smoothing over a limited range of frequencies.

The CCF algorithm is inherently monaural. Since we are using binaural signals, an intermediate step with ITD-based processing is performed using the PDCW algorithm that was discussed in Section 2.5.2. For reverberant input signals it is useful to first perform steady-state suppression to help reduce the effect of reverberation. The SSF algorithm is applied initially to both microphone signals to achieve this. The SSF algorithm is described in detail in Section 2.5.1.

6.1 Structure of the CCF algorithm

A block diagram describing CCF processing is shown in Figure 6.1.

6.1.1 Bandpass filtering

The CCF algorithm roughly mimics the manner in which speech is believed to be processed in the human auditory system. The peripheral auditory system is modeled by a bank of bandpass filters. We use a modified zero-phase implementation of the gammatone filters in Slaney's Auditory Toolbox [30]. The center frequencies of the filters are linearly spaced according to the ERB scale [16]. Gammatone filters are used because they

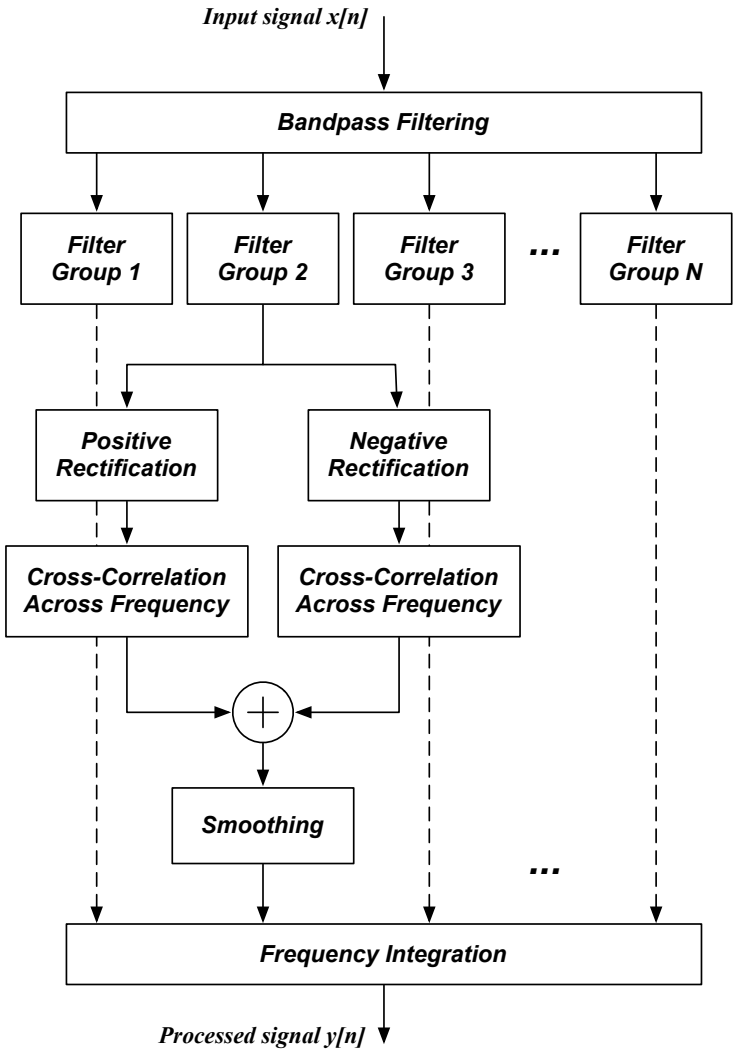


Figure 6.1: Block diagram describing the CCF algorithm.

approximate the frequency response of the peripheral auditory system. The frequency response of gammatone filters used is shown in Figure 6.2.

The impulse responses are obtained by computing the autocorrelation function of the original gammatone filters, which are adjusted to compensate roughly for the reduction in bandwidth produced by squaring the magnitude of the frequency response when performing the autocorrelation operation. This has been described in Section 4.2.1. Thus the effective filtering operation has the impulse response of

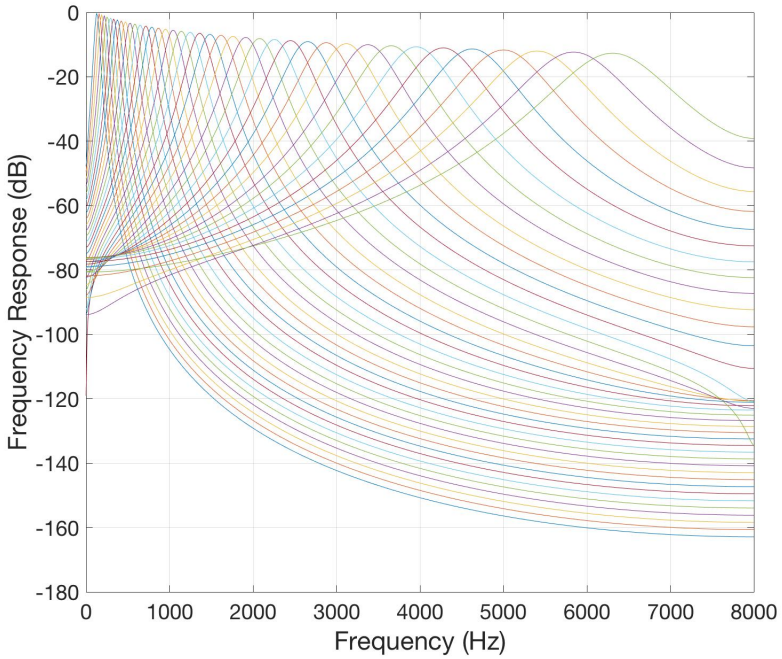


Figure 6.2: Frequency response of gammatone filters used in this study. This was generated using Malcolm Slaney's Auditory Toolbox.

$$h_e(n) = h(n) * h(-n) \tag{6.1}$$

where $h(n)$ is the original gammatone filter impulse response and $h_e(n)$ is the effective impulse response.

6.1.2 Satellite filters

Leveraging coherence in frequency requires that the outputs of a band of frequencies be considered. For the CCF algorithm, we look at a range of secondary channels on either side in frequency of the center frequency of the bandpass filters discussed in the previous section. Thus, for each of the primary gammatone filters used in conventional auditory processing, a secondary set of satellite filters is designed. The total span of these satellite

filters determines the range of frequencies over which CCF will be performed.

In other words, a total of N groups of bandpass filters is employed, each with one "center" band and $m/2$ satellite bands on either side of the center band in frequency, where m represents the total number of satellite bands. Since the satellite bands are symmetric about the center band, m is always even. These N filter groups are denoted by "Filter Group 1", "Filter Group 2" ...."Filter Group N" in Figure 6.1. Each of these filter groups consists of one center band and the corresponding satellite bands. The center frequency of the l^{th} pair of satellite filters on each side of the filter group center band is given by,

$$CB \pm s \times \alpha^{\frac{m}{2}+1-l}, \quad 1 \leq l \leq m/2 \tag{6.2}$$

where CB is the center band frequency for a given filter group, s is a parameter that determines the span of the frequencies on either side of the center band frequency and α is a parameter that controls the spacing between the satellite filters.

In this study, α was set to 0.7 which produces more closely spaced satellite filters closer to the center band and wider spacing away from the center band. This is physiologically motivated since it models the basilar membrane response. N was set to 20 and m was set to 6. The span parameter s was set to 2500 Hz. The values for the parameters mentioned above were determined experimentally.

Given the input signal $x[n]$, the filter outputs for a given filter group are given by

$$x_{kp}[n] = x[n] * h_{kp}[n] \tag{6.3}$$

where $x_{kp}[n]$ is the filter output of the k^{th} band of the p^{th} filter group, with $x[n]$ as input. Here k ranges from 1 to $m+1$ (comprising of m satellite bands and 1 center band) and p ranges from 1 to N.

6.1.3 Auditory-nerve-based processing

Bandpass filtering is followed by a rough model of auditory-nerve processing, which includes half-wave rectification of the filter outputs. Following earlier work in "polyaural" processing with multiple microphones [46], the filter outputs are also negated and similarly half-wave rectified. While this component of the processing is non-physiological, it enables the entire signal to be reconstructed, including positive and negative portions.

$$\begin{aligned} x_{+kp}[n] &= max(0, x_{kp}[n]) \\ x_{-kp}[n] &= max(0, -x_{kp}[n]) \end{aligned} \tag{6.4}$$

6.1.4 Cross-Correlation across frequency channels

Cross-correlation across frequency is then computed within each individual filter group.

$$\begin{aligned} X_{fcorr+p}[n] &= \left(\prod_{k=1}^{m+1} x_{+kp}[n] \right)^{\frac{1}{m+1}} \\ X_{fcorr-p}[n] &= \left(\prod_{k=1}^{m+1} x_{-kp}[n] \right)^{\frac{1}{m+1}} \end{aligned} \tag{6.5}$$

where $x_{+kp}[n]$ and $x_{-kp}[n]$ are the positive and negative half-wave-rectified portions of the signals $x_{kp}[n]$, and $X_{fcorr+p}[n]$ and $X_{fcorr-p}[n]$ denote the cross-correlation across frequency of $x_{+kp}[n]$ and $x_{-kp}[n]$ for the p^{th} filter group.

$X_{fcorr+p}[n]$ is combined with $-X_{fcorr-p}[n]$ to produce the complete cross-correlation

across frequency for the p^{th} filter group, $X_{fcorr_p}[n]$:

$$X_{fcorr_p}[n] = X_{fcorr+p}[n] + (-X_{fcorr-p}[n]) \tag{6.6}$$

In order to limit any distortion that may have taken place, the signal is bandpass filtered again to achieve smoothing. To resynthesize speech, all the filter groups are then combined.

$$y[n] = \sum_{p=1}^{N} \tilde{X}_{fcorr_p}[n] \tag{6.7}$$

where $\tilde{X}_{fcorr_p}[n]$ is the smoothed version of $X_{fcorr_p}[n]$.

6.2 Experimental Results

6.2.1 Experiments using simulated data

ASR performance using the CCF algorithm was tested using several combinations with algorithms for steady-state suppression and ITD-based analysis. An overall block diagram depicting the combinations used is shown in Figure 6.3. Preliminary results were obtained using the RM1 database using the Sphinx speech recognition system. Reverberation times of 0, 0.5 and 1 s were tested as part of the preliminary experiments. The data were simulated and an interfering talker was mixed in at 0, 10 and 20 dB SIR. Experiments were also conducted in the absence of an interfering talker. These results are tabulated in Table 6.1 and plotted in Figure 6.4.

Consider first the performance of the older compensation algorithms, PDCW and SSF, as described in Table 6.1. We note that PDCW provides excellent compensation for noise

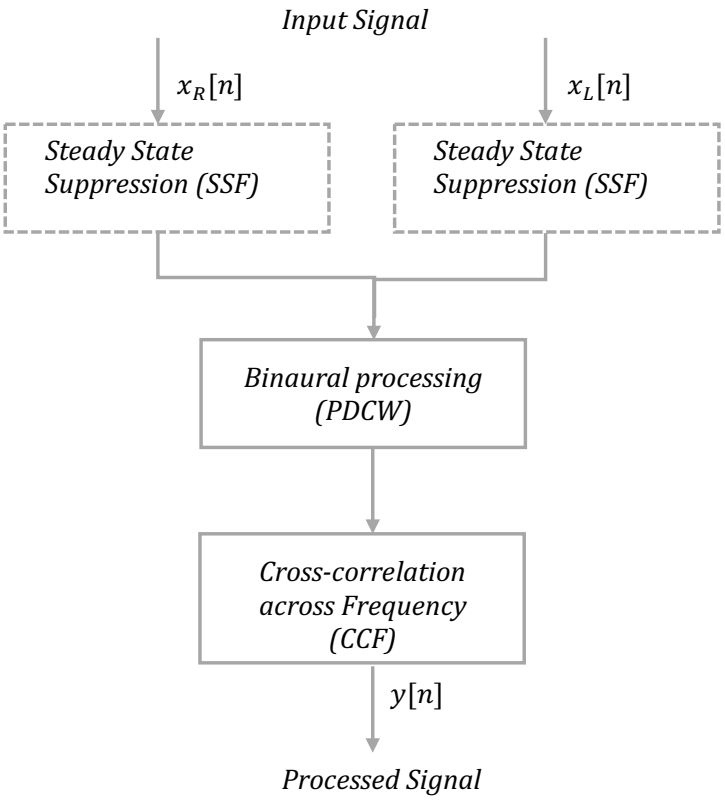


Figure 6.3: Block diagram describing the overall combination of algorithms used in conjunction with the CCF algorithm.

WER for $RT_{60} = 0$	RM1			
	0 dB	10 dB	20 dB	∞ dB
Delay and Sum	80.78%	32.01%	12.72%	6.54%
PDCW	23.01%	11.48%	8.15%	6.51%
PDCW+CCF	18.19%	11.48%	8.49%	7.48%
SSF	80.34%	31.31%	12.99%	6.82%
SSF+PDCW	25.43%	11.27%	7.78%	6.87%
SSF+PDCW+CCF	20.98%	12.21%	9.37%	8.51%

WER for $RT_{60} = 0.5s$	RM1			
	0 dB	10 dB	20 dB	Clean
Delay and Sum	95.95%	85.96%	66.44%	56.92%
PDCW	95.36%	86.64%	73.31%	66.63%
PDCW+CCF	94.56%	82.14%	68.53%	63.75%
SSF	97.14%	63.93%	35.03%	25.97%
SSF+PDCW	92.52%	61.64%	39.42%	32.27%
SSF+PDCW+CCF	84.65%	48.77%	32.53%	26.15%

WER for $RT_{60} = 1s$	RM1			
	0 dB	10 dB	20 dB	Clean
Delay and Sum	96.04%	92.5%	86.12%	82.52%
PDCW	96.08%	93.32%	89.08%	85.54%
PDCW+CCF	96.79%	93.84%	87.27%	84.18%
SSF	96.51%	78.96%	59.1%	52.17%
SSF+PDCW	94.75%	79.02%	63.63%	57.65%
SSF+PDCW+CCF	92.59%	68.2%	53.27%	46.78%

Table 6.1: Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interference Ratio for reverberation times of 0, 0.5 and 1 s for the RM1 database using the CMU Sphinx speech recognition engine using clean training data (Lowest WER for each condition highlighted)

WER for $RT_{60} = 0$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	43.43%	19.75%	5.25%
PDCW	10.48%	6.86%	5.4%
PDCW+CCF	11.9%	8.91%	7.85%
SSF	47.56%	20.94%	8.13%
SSF+PDCW	11.68%	8.74%	8.14%
SSF+PDCW+CCF	14.03%	10.52%	9.9%

WER for $RT_{60} = 0.2s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	54.6%	25.09%	13.66%
PDCW	58.4%	37.18%	30.6%
SSF	54.96%	23.71%	12.03%
SSF+PDCW	34.73%	19.13%	15.75%
SSF+PDCW+CCF	36.61%	20.53%	16.36%

WER for $RT_{60} = 0.4s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	82.5%	62.38%	52.74%
PDCW	91.8%	86.03%	81.06%
SSF	65.93%	34.69%	25.65%
SSF+PDCW	61.55%	39.9%	34.07%
SSF+PDCW+CCF	56.88%	37.4%	29.14%

WER for $RT_{60} = 0.6s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	91.86%	84.64%	81.49%
PDCW	94.51%	93.09%	91.01%
SSF	77.04%	53.84%	47.23%
SSF+PDCW	78.74%	62.66%	59.09%
SSF+PDCW+CCF	69.46%	51.37%	44.87%

Table 6.2: Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interference Ratio for reverberation times of 0, 0.2, 0.4 and 0.6 s for the WSJ database using GMM-based models trained using the Kaldi speech recognition toolkit using clean training data (Lowest WER for each condition highlighted)

WER for $RT_{60} = 0.8s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	94.81%	91.8%	88.74%
PDCW	95.33%	94.84%	93.2%
SSF	85.75%	68.17%	64.47%
SSF+PDCW	86.96%	76.13%	72.65%
SSF+PDCW+CCF	78.31%	64.09%	58.7%

WER for $RT_{60} = 1s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	95.11%	93.44%	90.94%
PDCW	95.55%	95.05%	94.06%
SSF	89.63%	77.97%	74.99%
SSF+PDCW	90.9%	83.73%	82.31%
SSF+PDCW+CCF	84.83%	72.91%	67.79%

Table 6.3: Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interference Ratio for reverberation times of 0.8 and 1 s for the WSJ database using GMM-based models trained using the Kaldi speech recognition toolkit using clean training data(Lowest WER for each condition highlighted)

in the absence of reverberation, but PDCW becomes less effective as the RT_{60} is increased from 0 to 1 seconds. SSF, in contrast, provides a good improvement in recognition accuracy in the presence of reverberation but its effectiveness is limited by the presence of interfering noise sources. Adding CCF to PDCW and SSF provides an even further improvement in WER, especially at low and moderate Signal-to-Interference Ratios (SIRs).

In the absence of reverberation, the PDCW algorithm provides signal separation as described in Chapter 2. The addition of CCF provides significant improvement of 20% relative in WER at 0 dB SIR. However, as the conditions get cleaner, applying CCF no longer leads to any improvements and in fact, the addition of CCF leads to much higher WER in the absence of reverberation and interfering talkers.

Some form of steady state suppression as performed by the SSF algorithm is required to achieve improvements in ASR in reverberant environments. In the presence of reverberation, the contribution of PDCW to ASR improvement is limited. However, in combination

WER for $RT_{60} = 0$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	30.77%	9.73%	3.16%
PDCW	6.52%	4.39%	3.19%
PDCW+CCF	8.61%	5.83%	4.63%
SSF	37.19%	11.34%	4.09%
SSF+PDCW	7.23%	5.34%	4.3%
SSF+PDCW+CCF	8.8%	6.87%	5.7%

WER for $RT_{60} = 0.2s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	47.88%	21.07%	10.42%
PDCW	47.66%	31.01%	25.44%
SSF	45.32%	16.08%	7.88%
SSF+PDCW	25.67%	14.46%	12.72%
SSF+PDCW+CCF	28.84%	16.03%	12.55%

WER for $RT_{60} = 0.4s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	75%	51.49%	41.01%
PDCW	83.41%	70.73%	63.4%
SSF	56.38%	26.1%	18.94%
SSF+PDCW	50.85%	31.23%	27.24%
SSF+PDCW+CCF	51.62%	32.79%	23.82%

WER for $RT_{60} = 0.6s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	87.56%	73.12%	65.37%
PDCW	92.12%	85.3%	81.8%
SSF	68.17%	42.54%	35.37%
SSF+PDCW	68.13%	47.36%	43.41%
SSF+PDCW+CCF	65.42%	46.12%	38.18%

Table 6.4: Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interference Ratio for reverberation times of 0, 0.2, 0.4 and 0.6 s for the WSJ database using DNN-based models trained using the Kaldi speech recognition toolkit using clean training data(Lowest WER for each condition highlighted)

WER for $RT_{60} = 0.8s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	90.86%	82.63%	76.82%
PDCW	93.31%	89.78%	86.83%
SSF	76.42%	54.98%	48.42%
SSF+PDCW	77.15%	63.05%	58.3%
SSF+PDCW+CCF	74.03%	57.95%	50.83%

WER for $RT_{60} = 1s$	WSJ		
	10 dB	20 dB	∞ dB
Delay and Sum	93.27%	86.34%	83.62%
PDCW	93.91%	90.77%	88.6%
SSF	83.41%	65.44%	60.58%
SSF+PDCW	84.85%	70.8%	66.15%
SSF+PDCW+CCF	81.45%	66.78%	59.33%

Table 6.5: Comparison of algorithms with respect to Word Error Rate as a function of Signal-to-Interference Ratio for reverberation times of 0.8 and 1 s for the WSJ database using DNN-based models trained using the Kaldi speech recognition toolkit using clean training data(Lowest WER for each condition highlighted)

with SSF and CCF, the improvements are significant. This is especially the case at moderate SIRs. The use of SSF+PDCW+CCF provides a relative improvement of nearly 21% at 10 dB compared to using SSF+PDCW for the 0.5 s reverberation-time case for RM1. These trends are quite consistent and hold even at the reverberation time of 1 s.

The experiments described above were repeated using the Kaldi speech recognition toolkit [38] and the WSJ database [39]. Experiments were conducted for reverberation times of 0.2 s to 1 s in steps of 0.2 s. These experiments were conducted using simulated data. An interfering talker was mixed in at 10 dB and 20 dB. Experiments were also conducted in the absence of an interfering talker. These experiments were conducted using both a GMM-based model and a DNN-based acoustical model trained using Kaldi. The DNN-based model used alignments generated using the GMM-based model. The DNN-based model has 2 hidden layers. Clean speech was used for training. Training data underwent SSF or CCF processing or both if that was part the test condition. Results

obtained from the GMM-based model are tabulated in Table 6.2 and 6.3. Some of the results specifically pertaining to the effect of the CCF algorithm are plotted in Figures 6.5 and 6.6.

As seen in Tables 6.2 and 6.3, the addition of the CCF algorithm doesn't help much at lower reverberation times. In fact, at reverberations times of 0 and 0.2 s, the baseline PDCW and SSF systems and their combination gives the best performance in terms of WER. As discussed in Chapter 2, the PDCW algorithm leads to significant gains in the absence of reverberation. Because of this, at the low reverberation time of 0.2 s, the combination of SSF+PDCW seems to give very good gains in WER. As the reverberation time increases, the effect of CCF becomes increasingly significant. At reverberation times of 0.6 s and higher, the combination of SSF+PDCW+CCF consistently has the lowest error.

Using DNN-based acoustical models gives slightly different results. As was the case with the GMM-based models, in the absence of reverberation, the addition of CCF does not lead to any improvements. The trends remain more or less the same as the GMM-based model at 0.2 and 0.4 s. However, at higher reverberation times, while the addition of the CCF algorithm does do better than using SSF+PDCW alone, sometimes the SSF algorithm by itself outperforms SSF+PDCW+CCF. In the presence of high reverberation and lower SIR, the addition of the CCF algorithm does lead to the best performing system in terms of WER. However, the baseline SSF system performs better at higher SIRs. However, the gap between the WER for the SSF system alone compared to SSF+PDCW+CCF does decrease as the reverberation time increases. While only reverberation times of up to 1 s have been tested here, it seems likely that the combination system of SSF+PDCW+CCF would lead to the lowest WERs for higher reverberation times.

The addition of the CCF algorithm to SSF and PDCW seems to give maximum gains in conditions where the mismatch between the training and test data are large. For this reason, multi-style training does not lead to any gains in WER and in fact, leads to poorer performance in terms of WER. Multi-style training results have not been reported.

6.2.2 Experiments using real reverberant data

In order to determine if the results obtained using simulated data can be generalized, experiments were also conducted using real reverberant data. The REVERB challenge database was used for this purpose [47]. The real reverberant data in the REVERB challenge consisted of utterances from the MC-WSJ-AV corpus which has utterances spoken in a noisy and reverberant room. Data collection took place in a room which had a reverberation time of about 0.7 s. The room contained two eight-element circular microphone arrays using which, the data were collected. The speakers were reading out sentences from the WSJCAM0 corpus [48]. While a number of different recording conditions existed, we used the first channel from one of the microphone arrays recorded in the stationary condition. The stationary condition referred to the condition where a speaker was asked to read a sentence from the WSJCAM0 corpus while stationary in the recording room. A triphone acoustical model which uses Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature transforms was trained using Kaldi. Clean training data were used for acoustical model training. The training data underwent processing that was identical to the test condition. The language model used was built using the WSJCAM0 database with a vocabulary of 5k words. In order to reduce the effect of the reverberant environment on the speech signals, SSF processing was performed as an initial step. The results obtained are tabulated in Table 6.6 and plotted in Figure 6.9.

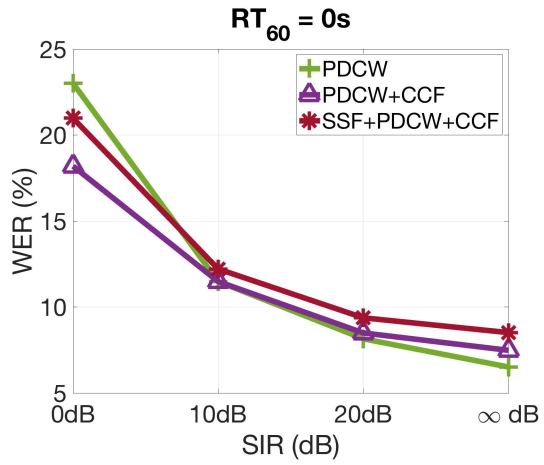
As seen in Table 6.6 and Figure 6.9, the combination of SSF and CCF leads to large gains in recognition accuracy compared to using SSF alone. The relative improvement is over 18%, highlighting the efficacy of the CCF algorithm.

Algorithm	WER(%)
Unprocessed	89.94%
SSF	59.89%
SSF+CCF	48.68%

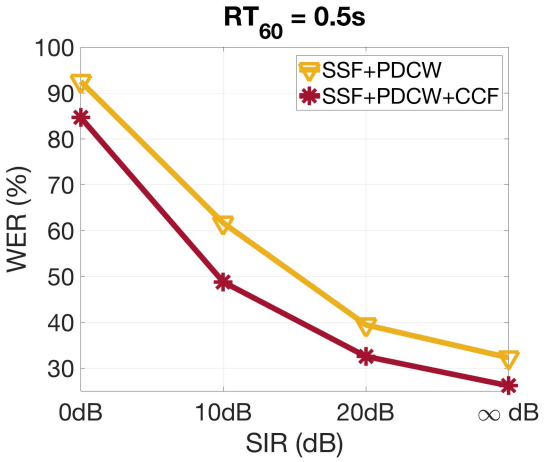
Table 6.6: Word Error Rate for algorithms tested using the REVERB challenge dataset. Only results using real reverberant data are reported here.

6.3 Conclusions

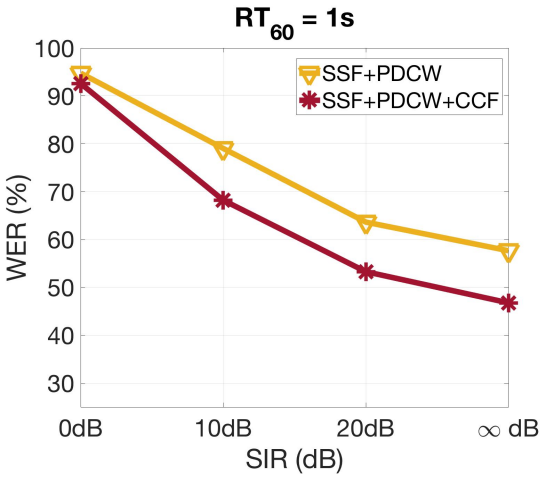
In this chapter we introduce a technique that uses coherence in the frequency domain to mitigate the effects of reverberation and noise. The CCF algorithm effectively boosts regions of coherence in frequency, the underlying assumption being that sounds coming from the same source will exhibit coherence in frequency. The CCF algorithm has been used in conjunction with the SSF and PDCW algorithms. This combination has been shown to lead to better performance in terms of WER under some conditions. In the presence of moderate to high reverberation, the improvements provided by the use of the CCF algorithms are observed primarily while using GMM-based models. These improvements diminish somewhat when DNN-based models are employed. Improvements using the CCF algorithm were also demonstrated with the use of real data.



(a)

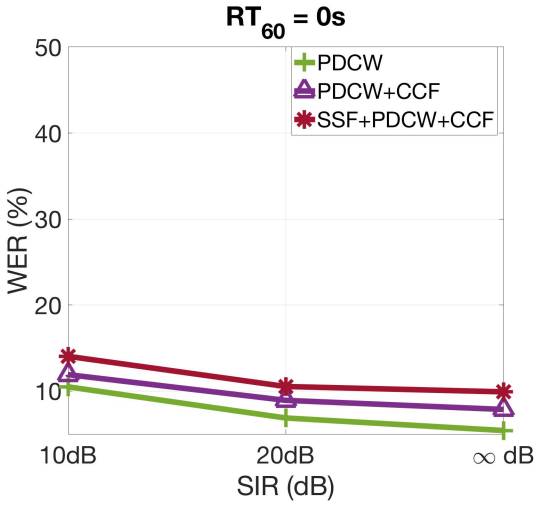


(b)

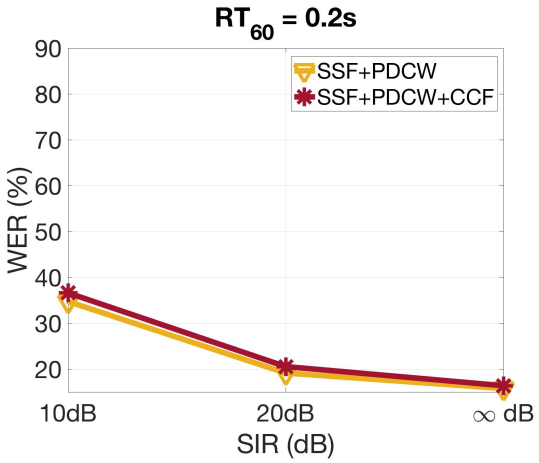


(c)

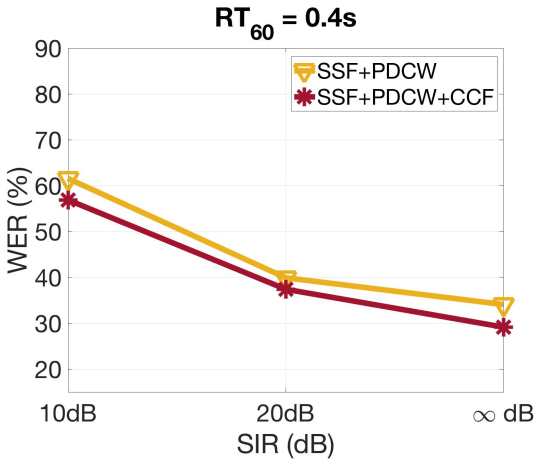
Figure 6.4: Word Error Rate evaluated using the CMU Sphinx speech recognition engine using clean training data for the RM1 database as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times (a) 0 s, (b) 0.5 s, (c) 1 s.



(a)

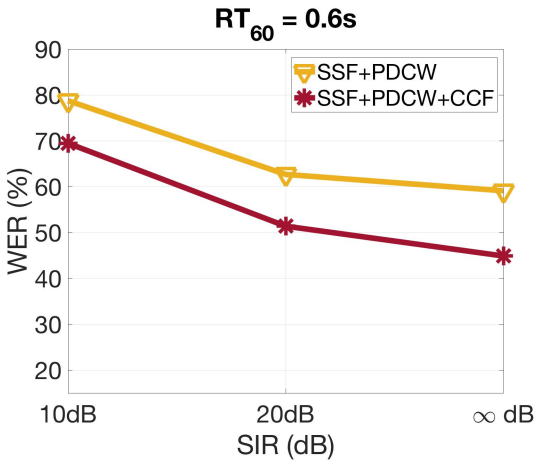


(b)

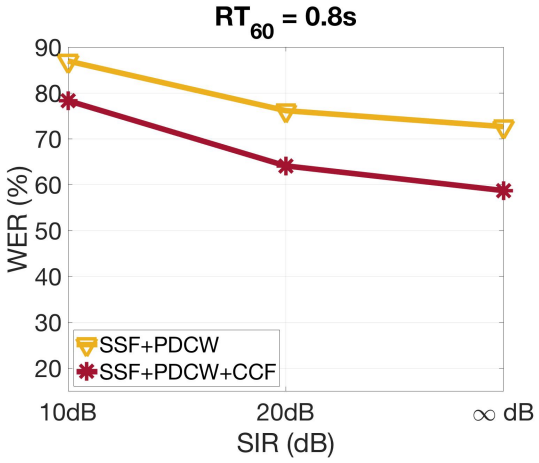


(c)

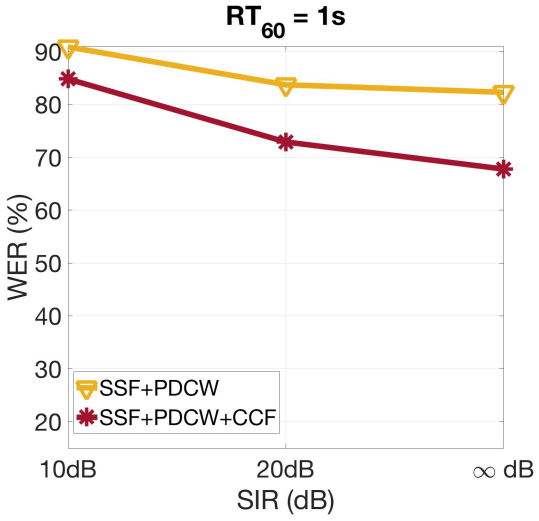
Figure 6.5: Word Error Rate evaluated using GMM-based models trained with the Kaldi speech recognition toolkit using clean training data for the WSJ database as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times (a) 0 s, (b) 0.2 s, (c) 0.4 s.



(a)

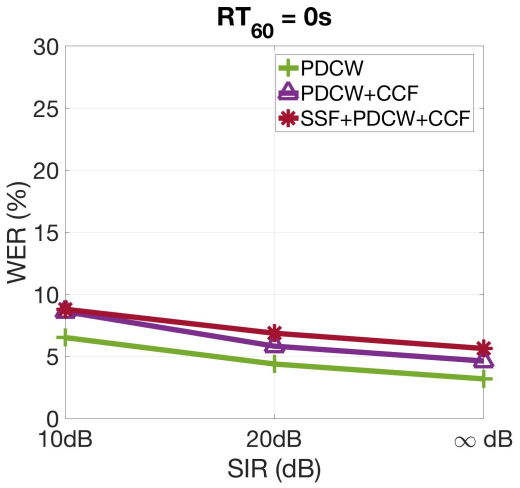


(b)

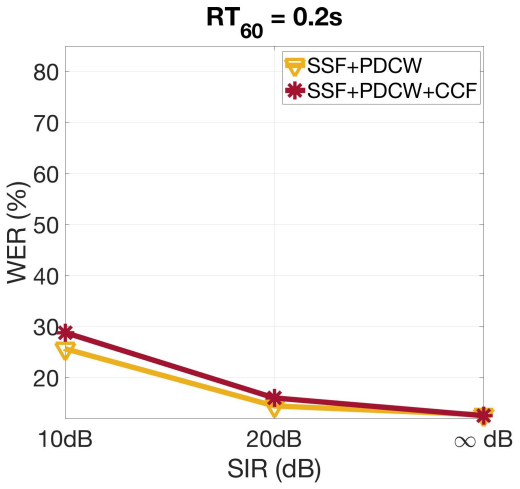


(c)

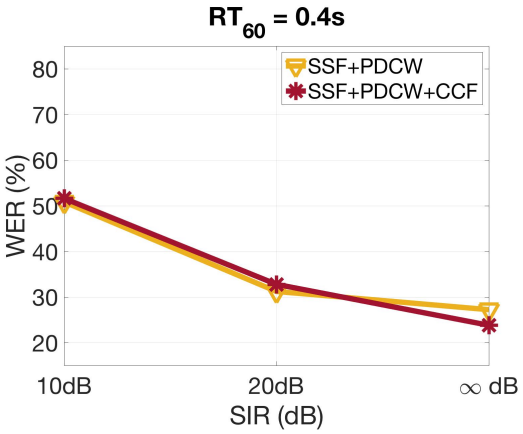
Figure 6.6: Word Error Rate evaluated using GMM-based models trained with the Kaldi speech recognition toolkit using clean training data for the WSJ database as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times (a) 0.6 s, (b) 0.8 s, (c) 1 s.



(a)

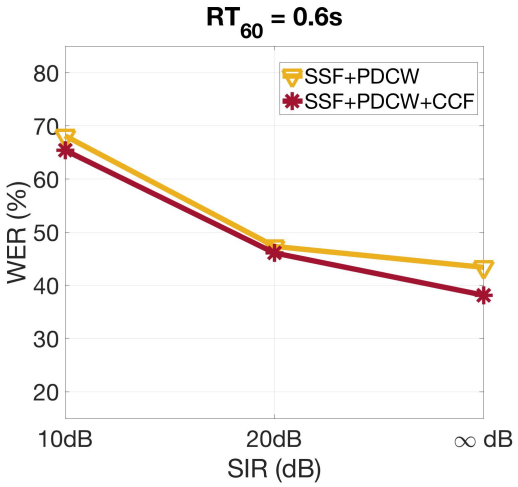


(b)

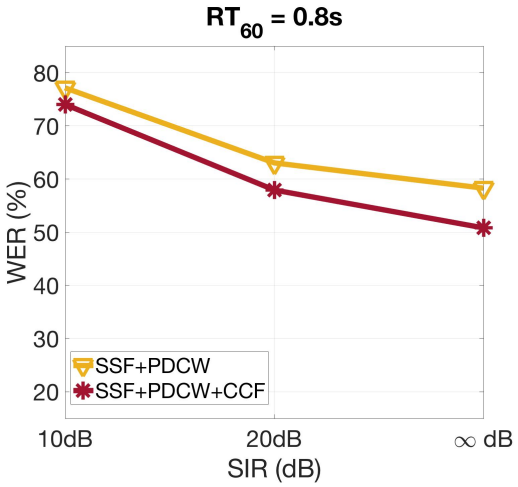


(c)

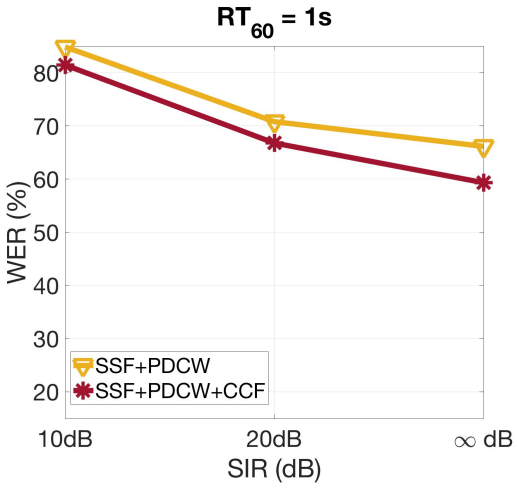
Figure 6.7: Word Error Rate evaluated using DNN-based models trained with the Kaldi speech recognition toolkit using clean training data for the WSJ database as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times (a) 0 s, (b) 0.2 s, (c) 0.4 s.



(a)



(b)



(c)

Figure 6.8: Word Error Rate evaluated using DNN-based models trained with the Kaldi speech recognition toolkit using clean training data for the WSJ database as a function of Signal-to-Interference Ratio for an interfering signal located 45 degrees off axis at reverberation times (a) 0.6 s, (b) 0.8 s, (c) 1 s. 87

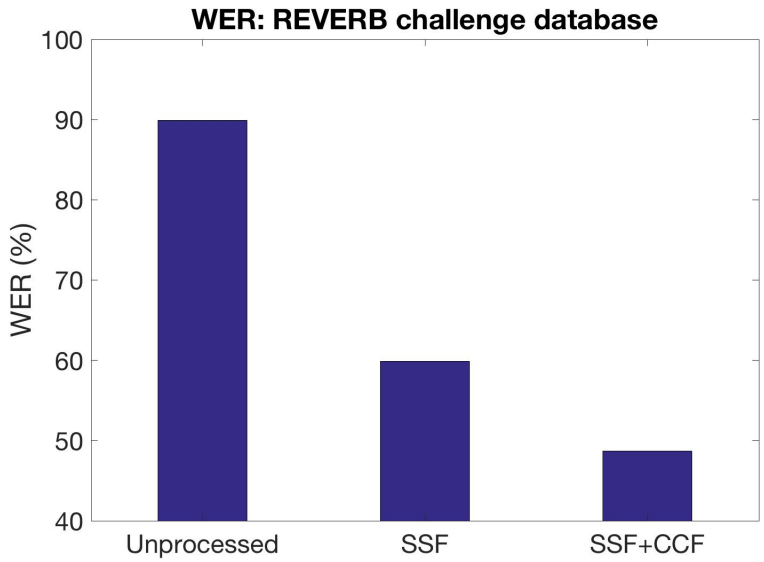


Figure 6.9: Results using the REVERB challenge database. The SSF and SSF+CCF algorithms are compared to the unprocessed signal.

CHAPTER 7

GENERAL DISCUSSION

In this chapter we briefly compare the algorithms introduced in Chapters 4, 5 and 6. Our specific aim is to achieve better ASR performance in the presence of reverberation and noise. For this purpose we examine the relative improvements seen in terms of ASR performance. The combination of SSF+ICW and SSF+CDRW are compared to (monaural) SSF processing. For the CCF algorithm, SSF+PDCW+CCF is compared to SSF+PDCW. Results using DNN-based models obtained by clean training as well as multi-style training are discussed here. The relative decreases in WER, averaged across all SIRs for different types of clean and multi-style training, are shown in Figures 7.1 and 7.2.

As seen in Figures 7.1 and 7.2, all improvements seen using the ICW, CDRW and CCF algorithm seem significant. Of the three techniques, the CDRW algorithm provides the best improvement in terms of ASR performance in the presence of reverberation and noise. The CDRW algorithm aims at suppressing regions of low coherence using the CDR metric. Thus, the gains using CDRW increase as the reverberation increases while using clean training, and the degree of mismatch between the training and test data also plays a significant role. As seen in Figure 7.2, while using multi-style training, the gains using the CDRW algorithm are greater at lower reverberation times. This probably has to do with the fact that most of the training data had moderate to high reverberation. Using clean

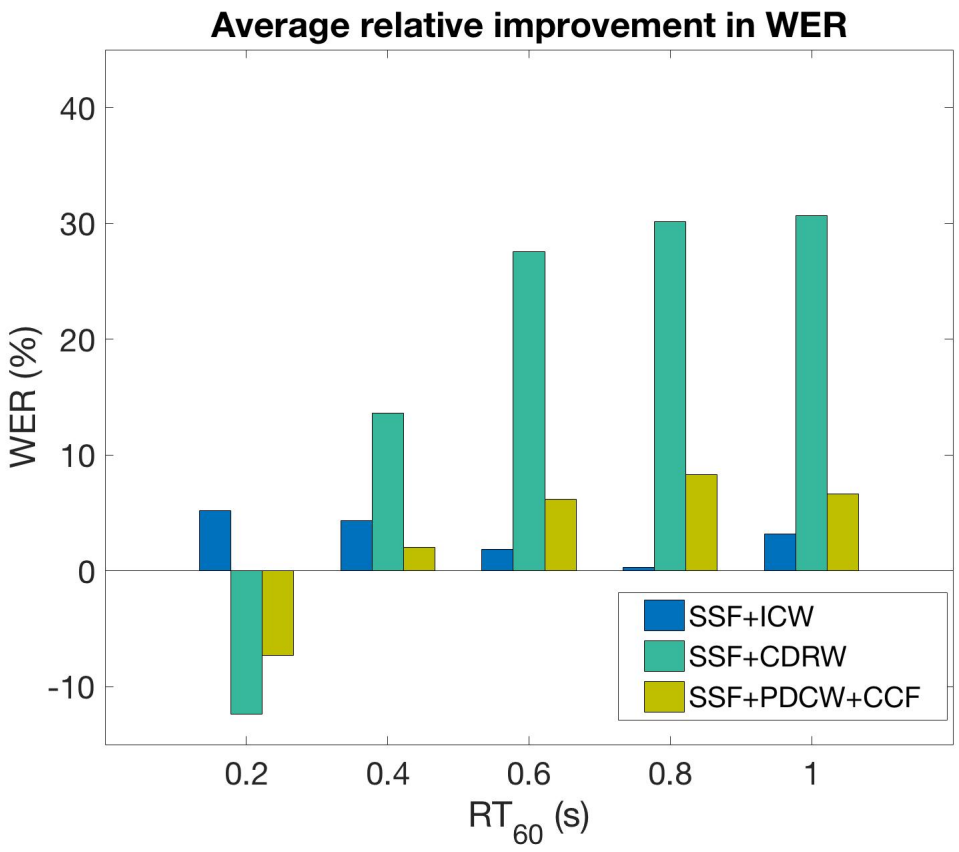


Figure 7.1: Comparison of the ICW, CDRW and CCF algorithms using clean training. Relative improvement in Word Error Rate averaged over all Signal-to-Interference Ratios, plotted as a function of reverberation time. Positive bars indicate better ASR performance. SSF (monaural) serves as baseline for SSF+ICW and SSF+CDRW while SSF+PDCW+CCF is compared to SSF+PDCW.

training, the CCF algorithm also gives greater gains at higher reverebration. Both the CDRW and CCF algorithms produce an increase in WER at low reverberation times using clean training as seen at $RT_{60} = 0.2$ s in Figure 7.1.

The Signal-to-Interference Ratio also plays a big part in ASR performance which may not be conveyed completely through Figures 7.1 and 7.2. For this reason, Figures 7.3, 7.4 and 7.5 are also provided here. Figure 7.3 shows the average improvements in WER observed using the SSF+ICW, SSF+CDRW and SSF+PDCW+CCF algorithms at 10 dB SIR and Figure 7.4 provides the same information for reverberated speech with no noise. Both use models trained using clean training data. Figure 7.5 shows the results using multi-

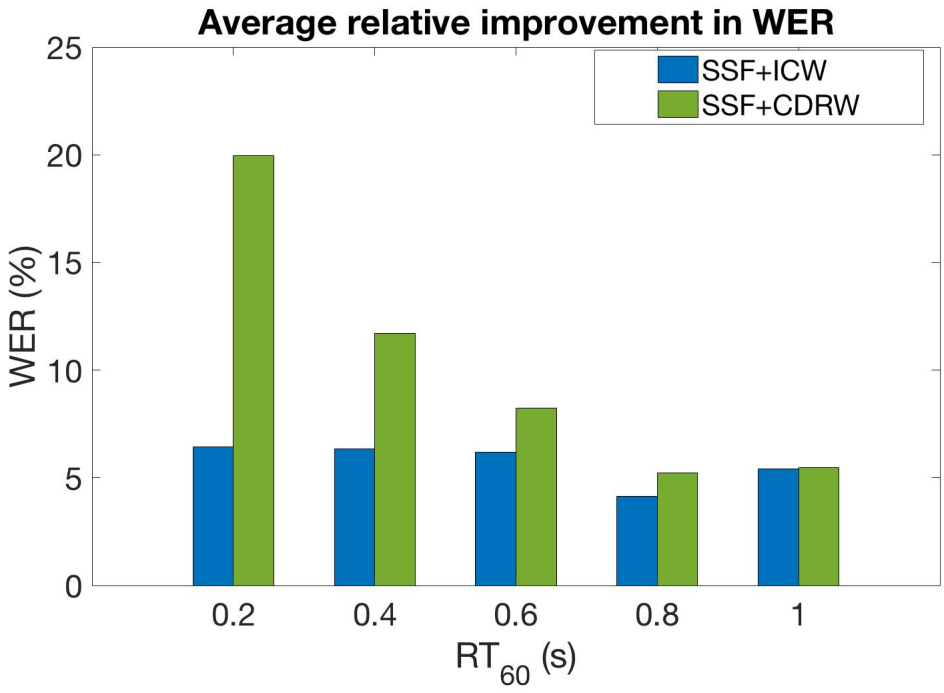


Figure 7.2: Comparison of the ICW and CDRW algorithms using multi-style training. Relative improvement in Word Error Rate averaged over all Signal-to-Interference Ratios, plotted as a function of reverberation time. Positive bars indicate better ASR performance. SSF (monaural) serves as baseline for SSF+ICW and SSF+CDRW while SSF+PDCW+CCF is compared to SSF+PDCW.

style training with the test data at 20 dB SIR.

As seen in Figures 7.3 and 7.4, the CDRW algorithm performs much better in the absence of an interfering talker. This is true for the CCF algorithm as well. Nevertheless, the performance of the CCF algorithm in the presence of an interfering talker at 10 dB is degraded significantly. In general, the CCF algorithm leads to greater improvements in WER in the presence of greater reverberation and at greater SIRs with SSF+PDCW as the baseline. The ICW algorithm, on the other hand, provides better improvements at low reverberation times. Improvements seen remain more or less similar across different values of SIR for the ICW algorithm.

Figure 7.5, which depicts data obtained using multi-style training, shows slightly different trends. In the case of the ICW algorithm, the improvements remain more or less constant across reverberation times at 20 dB SIR, with a slight dip at higher reverberation

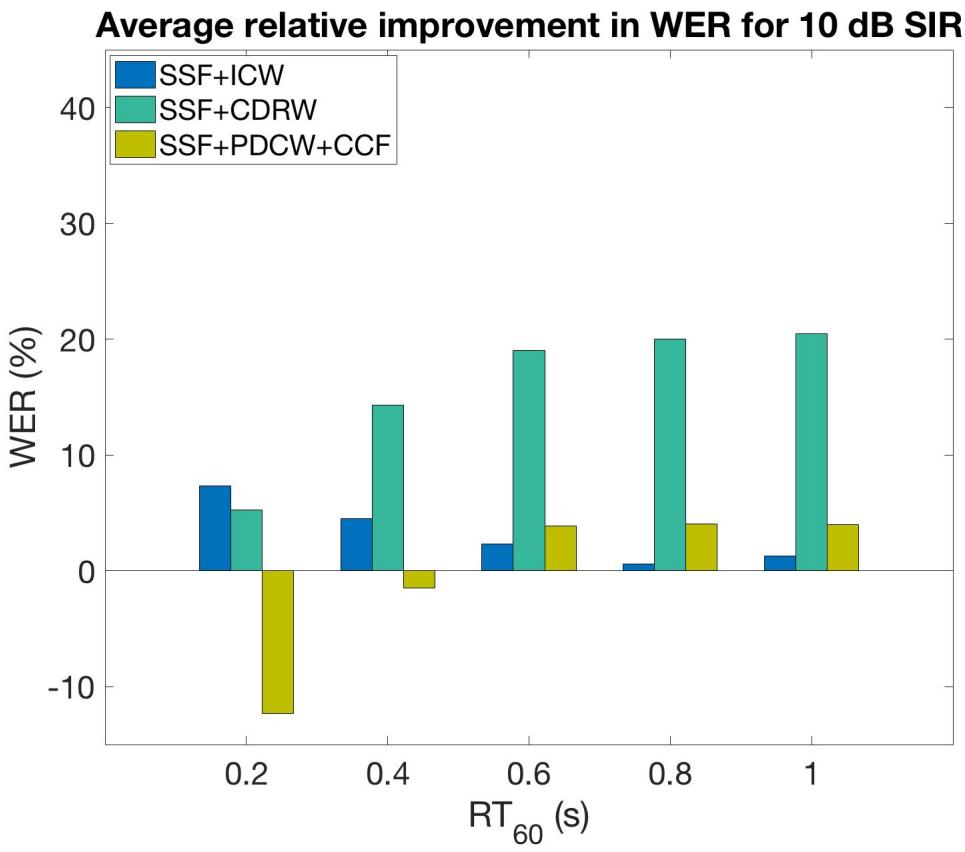


Figure 7.3: Comparison of the ICW, CDRW and CCF algorithms using clean training. Relative improvement in Word Error Rate at 10 dB SIR, plotted as a function of reverberation time. Positive bars indicate better ASR performance. Baselines are the same as Figures 7.1 and 7.2.

times. The results obtained using the CDRW algorithm show a much bigger difference. The addition of the CDRW algorithm to SSF leads to much greater improvements at lower reverberation times compared to using SSF alone. Even at reverberation time of 1 s, the relative improvement due to CDRW is over 5% relative. As mentioned before, the fact that the models were trained using moderate to high reverberation may have played a part in these trends.

Given the very significant improvements due to CDRW and CCF, we also attempted to combine both of them along with SSF. However, the combination of CDRW with CCF performed no better than the performance observed using SSF with CDRW alone. We have seen before that the CCF algorithm does not lead to improvements at low reverberation

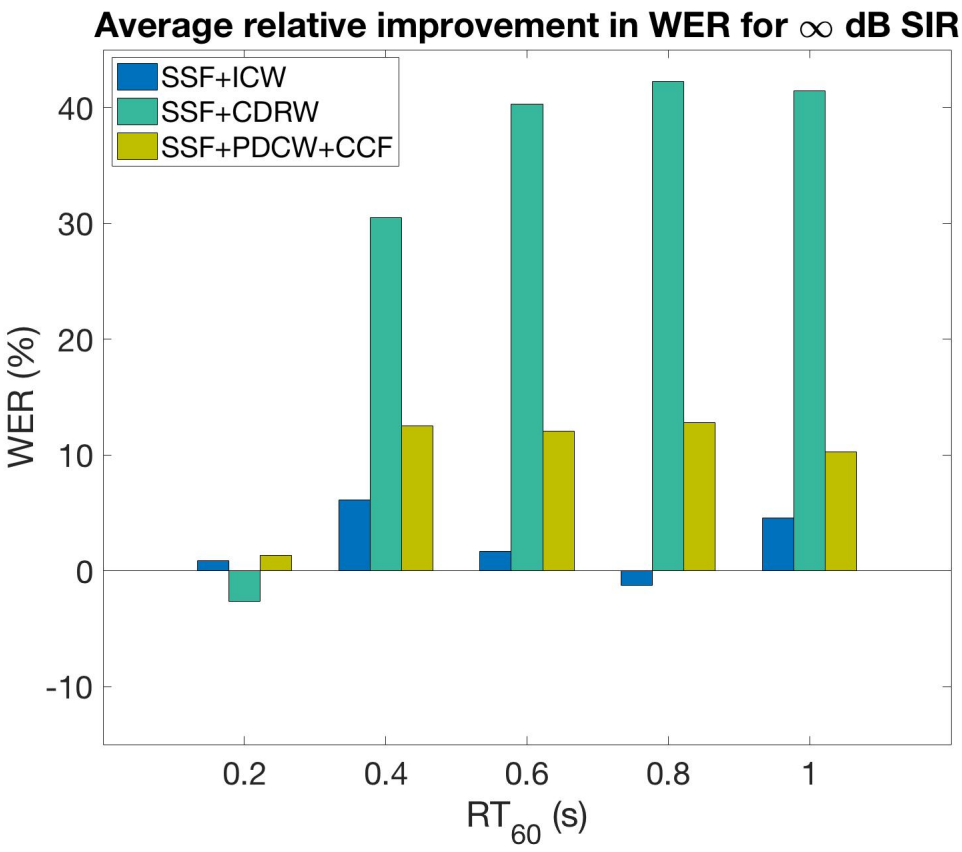


Figure 7.4: Comparison of the ICW, CDRW and CCF algorithms using clean training. Relative improvement in Word Error Rate at ∞ dB SIR, plotted as a function of reverberation time. Positive bars indicate better ASR performance. Baselines are the same as Figures 7.1 and 7.2.

times. It is possible that the addition of CCF to CDRW plus SSF does not provide further benefit because the impact of reverberation had already been suppressed sufficiently by the combination of SSF and CDRW.

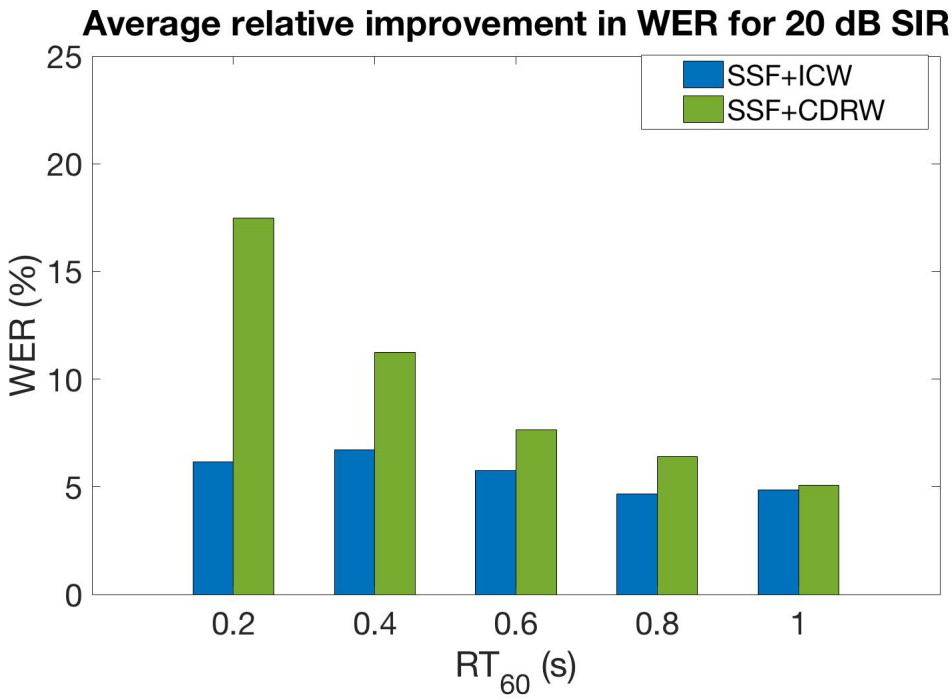


Figure 7.5: Comparison of the ICW and CDRW algorithms using multi-style training. Relative improvement in Word Error Rate at 20 dB SIR, plotted as a function of reverberation time. Positive bars indicate better ASR performance. Baselines are the same as Figures 7.1 and 7.2.

CHAPTER 8

SUMMARY AND CONCLUSIONS

In this thesis we address the problem of robust ASR in the presence of noise and reverberation. As listening environments go, this is a fairly common environment in which accurate ASR is necessary. This is especially true considering the widespread use of smart devices, most of which have a voice interface. In order to leverage our knowledge of the human auditory system, we develop techniques in this thesis that have their basis in how our ears achieve noise robustness in noisy or reverberant environments.

Three different techniques are discussed in this thesis to achieve better ASR accuracy in reverberant and noisy environments. This thesis deals specifically with binaural signals, and all three methods exploit coherence inherent in signals that originate from the same source in some way. In Chapter 4, we exploit coherence across the two microphones using the Interaural Cross-correlation-based Weighting (ICW). The ICW technique uses signal envelopes to perform cross-correlation across the two sensor signals. The human auditory system performs ITD analysis at higher frequencies using signal envelopes, and the ICW algorithm roughly follows this mechanism for binaural signal processing as well. The ICW algorithm lead to consistently better performance in preliminary studies using the RM1 database and the Sphinx speech recognition engine as well as in more detailed studies conducted using the WSJ database and the Kaldi speech recognition toolkit. In general,

the ICW algorithm provides greater improvements at lower levels of reverberation.

In Chapter 5, we exploit coherence across the two sensor signals once again. In contrast to the ICW algorithm, which made no assumptions about the listening environment, a model-based method is described in this chapter. Specifically, the Coherence-to-Diffuse Ratio-based Weighting (CDRW) method uses a model for coherence in a diffuse field versus in a coherent field to compute the Coherent-to-Diffuse Ratio (CDR) metric for each time-frequency bin of the STFT of the input signals. This is then used as a weight in order to suppress regions with low CDR. The CDRW algorithm was also tested in conditions similar to the ICW algorithm and it showed very significant gains (over 40% relative) especially for conditions of moderate reverberation using clean training. Using multi-style training, these trends were different with better performance at lower reverberation times. Across training styles, the CDRW algorithm leads to significant improvements.

Chapter 6 describes a technique that uses coherence in frequency of signals that originates from the same source. This method is called Cross-Correlation across Frequency (CCF) and it is roughly based on the concept of "straightness weighting," which hypothesizes that greater emphasis is given to ITDs that are consistent in frequency over some limited range in terms of auditory perception. Unlike the ICW and CDRW algorithms, the CCF algorithm is not binaural in nature. Therefore, an intermediate binaural processing step is required before the application of CCF. The CCF algorithm then performs cross-correlation across a small range of frequencies around a center band of interest. The application of CCF provides gains especially in the the presence of high reverberation using clean training. Multi-style training results for CCF were not obtained.

8.1 Future work

There exist several opportunities for further investigation in the techniques discussed as part of this thesis. Of all the algorithms discussed, the CDRW algorithm definitely provides the greatest improvement in error rate under many conditions. Nevertheless, the way the CDRW algorithm is currently implemented does not actively suppress interfering talkers from a known location. In order to this, we need to update the generalized complex coherence function and take into account every known source. This will most likely lead to better performance of CDRW processing even at low SIRs.

Several techniques discussed in this thesis produce a weight matrix or mask as their output. It is hypothesized that these masks boost the regions where the target signal is dominant and effectively attenuate portions of the signal dominated by noise. Using the various features we have at our disposal, including features based on the techniques discussed in Chapters 4, 5 and 6, it is possible to learn a mask that is as close as possible to the Ideal Ratio Mask (IRM). This now reduces to a supervised learning problem.

For some time now, the use of deep-learning techniques for supervised classification problems has been shown to be very effective [49, 50]. Neural networks can model complex non-linear relationships which makes them very useful for tasks like mask estimation using multiple features. For this reason, deep-learning-based approaches can be used towards developing IRMs from a suite of complementary features derived from binaural speech signals.

One possible technique involves the use of a Deep Neural Net (DNN) for mask estimation similar to those seen in [49, 50]. A feature set that is a combination of features obtained from SSF, CDRW, CCF, ICW and other techniques discussed in this thesis in conjunction with standard features such as PNCC, RASTA-PLP [51, 52] may be used.

REFERENCES

[1] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, IEEE, 2013, pp. 7398–7402.

[2] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, 2014, pp. 1759–1763.

[3] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[4] C. Kim and R. M. Stern, "Nonlinear enhancement of onset for robust speech recognition.," in *INTERSPEECH*, 2010, pp. 2058–2061.

[5] C. Kim, K. Kumar, and R. M. Stern, "Binaural sound source separation motivated by auditory processing," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, 2011, pp. 5072–5075.

[6] R. M. Stern, C. Kim, A. Moghimi, and A. Menon, "Binaural technology and automatic speech recognition," in *International Congress on Acoustics*, 2016.

[7] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 4, pp. 361–378, 2004.

[8] L. A. Jeffress, "A place theory of sound localization.," *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[9] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *Journal of the Acoustical Society of America*, vol. 35, no. 8, pp. 1206–1218, 1963.

[10] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination," *The Journal of the Acoustical Society of America*, vol. 54, no. 6, pp. 1458–1470, 1973.

[11] R. M. Stern, G. J. Brown, D. Wang, D Wang, and G. Brown, "Binaural sound localization," *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, pp. 147–185, 2006.

[12] W. Kock, "Binaural localization and masking," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 801–804, 1950.

[13] R. M. Stern and C. Trahiotis, "Models of binaural interaction," *Handbook of perception and cognition*, vol. 6, pp. 347–386, 1995.

[14] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[15] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," PhD thesis, Carnegie Mellon University, 2010.

[16] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[17] C. Kim and R. M. Stern, "Power function-based power distribution normalization algorithm for robust speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, IEEE, 2009, pp. 188–193.

[18] C. Kim, K. Kumar, B. Raj, and R. M. Stern, "Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain.," in *INTERSPEECH*, Citeseer, 2009, pp. 2495–2498.

[19] 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.

[20] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The darpa 1000-word resource management database for continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, IEEE, 1988, pp. 651–654.

[21] G. B. Henning, "Detectability of interaural delay in high-frequency complex waveforms," *The Journal of the Acoustical Society of America*, vol. 55, no. 1, pp. 84–90, 1974.

[22] J. E. Rose, N. B. Gross, C. D. Geisler, and J. E. Hind, "Some neural mechanisms in the inferior colliculus of the cat which may be relevant to localization of a sound source.," *Journal of Neurophysiology*, vol. 29, no. 2, pp. 288–314, 1966.

[23] A. Menon, C. Kim, and R. M. Stern, "Robust speech recognition based on binaural auditory processing," 2017.

[24] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization (tutorial reprint)," *Journal of the Audio Engineering Society*, vol. 21, no. 10, pp. 817–826, 1973.

[25] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[26] P. M. Zurek, "The precedence effect," in *Directional hearing*, Springer, 1987, pp. 85–105.

[27] K. D. Martin, "Echo suppression in a computational model of the precedence effect," in *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, IEEE, 1997, 4–pp.

[28] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals," *Journal of the Acoustical Society of America*, vol. 80, pp. 1608–1622, 1986.

[29] S. O. Sadjadi and J. H. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, 2011, pp. 5448–5451.

[30] M. Slaney, "Auditory toolbox version 2," *University of Purdue, https://engineering. purdue. edu/˜ malcolm/interval/1998-010*, 1998.

[31] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[32] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Signal Processing Conference, 2011 19th European*, IEEE, 2011, pp. 1347–1351.

[33] O. Thiergart, G. Del Galdo, and E. A. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, 2012, pp. 309–312.

[34] J. Allen, D. Berkley, and J Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[35] C. Zheng, A. Schwarz, W. Kellermann, and X. Li, "Binaural coherent-to-diffuse-ratio estimation for dereverberation using an itd model," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, IEEE, 2015, pp. 1048–1052.

[36] A. Westermann, J. M. Buchholz, and T. Dau, "Binaural dereverberation based on interaural coherence histograms a," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 2767–2777, 2013.

[37] E. A. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[39] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, 1992, pp. 357–362.

[40] B. M. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech," *The Journal of the Acoustical Society of America*, vol. 29, no. 9, pp. 973–987, 1957.

[41] R. M. Stern and C. Trahiotis, "Models of binaural interaction," *Handbook of perception and cognition*, vol. 6, pp. 347–386, 1995.

[42] H. S. Colburn and A. Kulkarni, "Models of sound localization," in *Sound source localization*, Springer, 2005, pp. 272–316.

[43] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, "Lateralization of complex binaural stimuli: A weighted image model," *Journal of the Acoustical Society of America*, vol. 84, pp. 156–165, 1988.

[44] R. M. Stern and C. Trahiotis, "The role of consistency of interuaral timing over frequency in binaural lateralization," in *Auditory physiology and perception*, Y. Cazals, K. Horner, and L. Demany, Eds., Pergamon Press, Oxford, 1992, pp. 547–554.

[45] ——, "Binaural mechanisms that emphasize consistent interaural timing information over frequency," in *Proceedings of the XI International Symposium on Hearing*, A. R.

Palmer, A. Rees, A. Q. Summerfield, and R. Meddis, Eds., Whurr Publishers, London, 1998.

[46] R. M. Stern, E. B. Gouvêa, and G. Thattai, "Polyaural array processing for automatic speech recognition in degraded environments," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[47] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *et al.*, "A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[48] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: A british english speech corpus for large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, IEEE, vol. 1, 1995, pp. 81–84.

[49] X. Zhang and D. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[50] Z.-Q. Wang and D. Wang, "Robust speech recognition from ratio masks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, IEEE, 2016, pp. 5720–5724.

[51] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, IEEE, vol. 1, 1992, pp. 121–124.

[52] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, 2012, pp. 4101–4104.