A Spectro-Temporal Framework for Compensation of Reverberation for Speech Recognition

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Kshitiz Kumar

B.Tech., Electrical Engineering, Indian Institute of Technology, Kharagpur

Carnegie Mellon University
Pittsburgh, PA

May, 2011

Dedicated to my parents

Laxman P. Agrawal and Satyam Devi,

they took undue effort in giving me the gift of education.

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my Ph.D. thesis advisor and the committee chair, Richard M. Stern for his unwavering interest and support for my work. Throughout my five-and-a-half years of graduate life at Carnegie Mellon University (CMU), I have seen him display impeccable integrity, exemplary professionalism, hard work and a huge passion for speech research and teaching. He was always for "fully-baked" problem solving, was very patient with research results and encouraged luxurious thinking. I feel extremely fortunate to have had his blessings in my thesis work.

Along with Rich, Rita Singh and Bhiksha Raj mentored and supported my crucial penultimate year at CMU. They were always available for long discussions and participated in my thesis committee as well. Rita offered numerous great ideas and immensely helped improve my thesis proposal and defense. Bhiksha identified many missing dots in my thesis and emphasized on creating a good story out of the thesis. Alex Acero, Paris Smaragdis, and Vijayakumar Bhagavatula constituted the rest of my thesis committee. I am very thankful for their time and constructive criticism of my thesis. I am also grateful to Tsuhan Chen for his initial support and mentorship.

I am also very thankful to my Robust Speech Group colleagues and friends Evandro Gouvea, Ziad Al-Bawab, Lingyun Gu, Yu-Hsiang Chiu, Chanwoo Kim, Amir Moghimi, and Mark Harvilla. They added new dimensions to my times at CMU. Evandro greatly helped me in my initial software-and-tools learning stage at CMU. My collaborations with Chanwoo were especially fruitful.

My times at CMU were further enriched by my association with Indian Graduate Students Organization (IGSA). At CMU, I had a pleasure to also collect many great friends in Puneet Aggarwal, Dhruv Batra, Andrew Gallaghar, Niti Garg, Dhiraj Goel, Ashfaque Habib, Sumit Jha, Soummya Kar, David Liu, Sharad Oberoi, Devi Parikh, Aabid Shariff, Prabhanshu Shekhar, and Abhinav Singhal. Each of them added uniquely their own color into my life at CMU.

Among the ECE administrative staff, I am especially thankful to Carol Patterson,

Lyz Knight and Elaine Lawrence. They were my extended family at CMU.

My undergraduate studies in Electrical Engineering at Indian Institute of Technology (IIT), Kharagpur served as the stepping-stones towards my Ph.D. Praveen and Shyamesh Kumar were my other friends who immensely helped in setting up a good foundation for a scholarly career.

Finally tributes are due for my family members. My parents Laxman P. Agrawal and Satyam Devi took undue effort for my education - they were my greatest motivators. My wife Neha showed a great confidence in me, this work would not have been possible without her love, support and encouragement that kept me going even when the going got tough. I am also very thankful for the support I got from her parents Om Prakash and Geeta Agrawal. My siblings Sandhya, Shashi, Pawan and Ravi Agrawal stood by me in all my good and bad times - and so were my grandparents Viswanath and Panna Agrawal, their blessings shall forever be with me.

This work was supported by the National Science Foundation (Grant IIS-10916918 and IIS-10916918) and the C. Stark Draper Laboratory.

ABSTRACT

The objective of this thesis is the development of signal processing and analysis techniques that would provide sharply improved speech recognition accuracy in highly reverberant environments. Speech is a natural medium of communication for humans, and in the last decade various speech technologies like automatic speech recognition (ASR), voice response systems etc. have considerably matured. The above systems rely on the clarity of the captured speech but many of the real-world environments include noise and reverberation that mitigate the system performance. The key focus of the thesis is on the robustness of ASR to reverberation.

In our work, we first provide a new framework to adequately and efficiently represent the problem of reverberation in speech feature domains. Although our framework incurs modeling approximation errors, we believe that it provides a good basis for developing reverberation compensation algorithms. Based on our framework, we successfully develop a number of dereverberation algorithms. The algorithms reduce the uncertainly involved in dereverberation tasks by using speech knowledge in terms of cepstral auto-correlation, cepstral distribution, and, non-negativity and sparsity of spectral values. We demonstrate the success of our algorithms on clean-training as well as matched-training.

Apart from dereverberation, we also provide an approach for noise robustness via a temporal-difference operation in the speech spectral domain. There, via a theoretical analysis, we predict an expected improvement in the SNR threshold shift for white-noise conditions. We also empirically quantify and study speech-feature level distortion with respect to speech-signal level additive noise.

Finally, we provide a new framework for a joint reverberation and noise representation and compensation. The new framework generalizes the spectral domain reverberation framework by incorporating an additive noise term. Working under the new framework, we combine our dereverberation and noise compensation approaches for better dereverberation as well as for the most challenging speech recognition task that includes both noise and reverberation components.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The objective of this thesis [1] is the development of signal processing and analysis techniques that would provide sharply improved speech recognition accuracy in highly reverberant environments. In the following section we provide a brief introduction to speech technology, discuss the issue robustness for speech technologies to noise and reverberation, understand and model the phenomenon of reverberation for speech recognition, and briefly introduce a conventional feature extraction procedure for automatic speech recognition (ASR). Finally, we present the thesis organization.

1.1 SPEECH TECHNOLOGY

Speech technologies have considerably matured in the last decade and serve as the basis for numerous speech-based applications. Some of the examples of key speech based technologies are Automatic speech recognition (ASR), speaker recognition (SRE), and speech translation (ST). Speech is the natural medium of communication for humans, and the growth of speech technologies have greatly advanced the human-computer interaction (HCI) by enabling the computer to "listen" and "talk". The applications of HCI have resulted in some of the key technologies of today such as interactive voice response (IVR), dictation systems, and voice-based command and control for robots, etc. The speech characteristics of a person are very unique to the person, and consequently the speech technologies have been greatly successful in speaker identification and verification tasks. Speech technologies have not only reduced the communication gap between humans and machines but also between humans speaking different languages through speech to speech translation. Speech technologies have indeed been one of the highly successful technologies of the current generation.

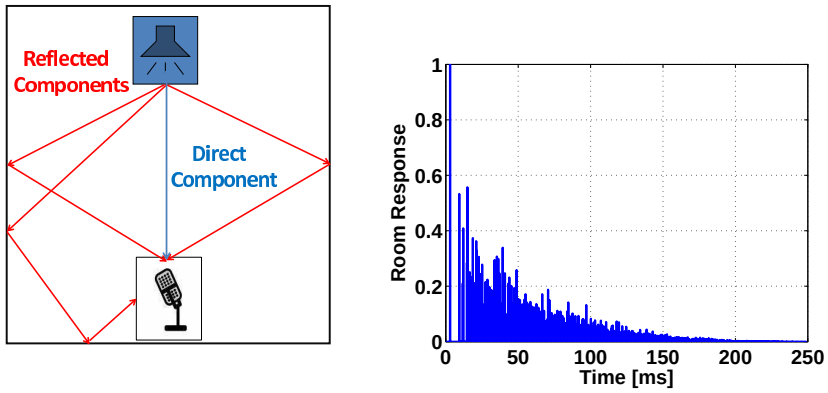
1.2 ENVIRONMENTAL ROBUSTNESS FOR SPEECH TECHNOLOGIES

Current state-of-the-art speech-based systems perform very well in the controlled environments where the speech signals are reasonably clean. But real-life conditions are far less controlled and include noise and reverberation in the environment. While human speech perception is remarkably robust to noise and reverberation, speech perception by machines is very sensitive to environmental conditions. This has considerably affected the widespread deployment of speech technologies in practice. The issue of environmental robustness has been the object of significant attention in the last decade. A number of algorithms have been successfully developed for robustness to noise but reverberation remains a challenging problem.

1.3 UNDERSTANDING REVERBERATION

Reverberation is an acoustic phenomenon in which a sound wave traveling in an enclosure is repeatedly reflected by the difference surfaces in the enclosure. Thus, reverberation lets the sound persist even after original sound is switched off. We further explain reverberation with respect to Fig. 1.1(a). There, the sound source radiates sound with a specified directivity pattern. Sound waves hitting an enclosure surface will be partly absorbed, partly transmitted and the rest attenuated and reflected back into the enclosure. The amount of sound absorption, transmission and reflection depends on the surface material and the sound frequency. The amount of energy in the reflected components will gradually diminish due to absorption by the surface materials. Reverberation for an enclosure is parameterized in terms of reverberation time (RT), which is the time taken for the signal power to decay by 60 dB from the instant the signal source is switched off. Thus, environments with greater RTs imply a longer persistence of the sound in the enclosure after its source is switched off.

Reverberation thus creates a collection of reflected and attenuated sounds in an enclosure. These reflected sounds interfere with and distort the original sound. A "listener" in the environment (see Fig. 1.1(a)) will hear the direct signal component as well as the reflected components. The impact of reverberation on human auditory perception depends upon the room RT. If the RT is small, the environment will rein-



(a) (b)

Fig. 1.1: *(a) Direct and reflected signal components, (b) Simulated room impulse response (RIR) for an environment with RT of 300 ms.*

force the sound which may enhance the sound perception. But if the room RT is large, a spoken syllable may persist for long and interfere with the spoken syllables in future [1]. This severely degrades speech perception [2–4]. Since reverberation is a collection of reflected and attenuated sounds, it is mathematically modeled as a linear system to represent the delayed and attenuated components of the sound. Fig. 1.2 presents a physical model of reverberation.

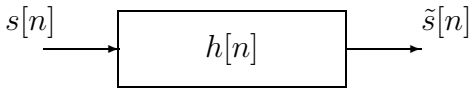


Fig. 1.2: *Modeling reverberation for a time-domain signal.*

The signal $s[n]$ in Fig. 1.2 represents a discrete-time clean speech signal, $\tilde{s}[n]$ represents the reverberated signal and $h[n]$ represents the reverberation filter mapping $s[n]$ to $\tilde{s}[n]$. The filter $h[n]$ represents the room impulse response (RIR) of the linear system from the source $s[n]$ to the microphone which receives the signal. The RIR will in general depend on the room characteristics including the geometry of the room, source and receiver locations, and absorption properties of the objects and walls in the

room. Thus, the characteristics of the filter $h[n]$ will change according to changes in the room characteristics. It is typically assumed that compared to changes in the speech spectral characteristics, the room spectral characteristics change slowly and that room characteristics are nearly stationary over a short time interval (1-2 s). Thus, the filter $h[n]$ is assumed to be time-invariant and overall it is a linear time-invariant (LTI) system.

We plot a simulated RIR for a room with RT of 300 ms in Fig. 1.1(b). There, we note that the reflected components are initially sparse but as reverberation gradually builds up, the rate of arrival of the reflected components increases very sharply. Finally, after about 300 ms the sound energy in the enclosure have decayed by 60 dB. The above RIR was simulated by image method proposed by Allen and Berkeley in [5]. We used the *RIR* program [6] to simulate image-method-based RIR. The simulation assumed a rectangular shoe-box room. We will be using [6] for a number of reverberation related experiments in later chapters. *Roomsim* [7] is an alternate RIR simulation tool that compared to [6] offers many more parameters in the simulation process.

1.4 IMPACT OF REVERBERATION ON ASR

Since the objective of the proposed work is robustness of ASR to reverberation, it is important to understand how reverberation affects ASR performance. As noted in Sec. 1.3, reverberation causes a sound to persist in a medium even after the sound source is switched off. Thus, reverberation will lead to temporal and spectral smearing of a signal [8–10], which will significantly distort the perceived sound. The distortion of the signal and hence its spectrum is very harmful for ASR, as ASR is essentially a pattern-matching algorithm based on the features derived from the signal spectral patterns. The distorted spectral patterns will not match well to the corresponding clean spectral patterns, resulting in degraded ASR performance [11]. We further illustrate the impact of reverberation on ASR in Fig. 1.3. There, we plot the spectrogram (see Sec. 1.5) corresponding to a typical clean speech signal in Fig. 1.3(a). Fig. 1.3(b) plots the spectrogram corresponding to reverberated speech with an RIR at RT of 300 ms in Fig. 1.1(b). The mismatch between the spectrograms in Fig. 1.3(a) and (b) is discernible and very harmful for the current ASR systems. We plot the word error rate (WER) for DARPA RM Database [12] [13] in the presence of reverberation in

Fig. 1.4, where we note that the ASR performance degrades rapidly in reverberation with the WER jumping from 6.7% for clean speech to 51% for RT of 300 ms.

Figure 1.4 is a plot of the ASR for mismatched training and testing conditions, where training was done on clean speech and testing was performed using different reverberation conditions. With an oracle knowledge of the test environment, we could also perform a matched training and testing, where training is also done on data from the test environment. It is worthwhile noting that even with matched training and testing the WER is significantly high. Specifically, for matched training the WER increases from 6.7% for clean conditions to 20% for RT of 500 ms. Thus, even matched training and testing, which incorporates oracle knowledge does not provide good ASR performance. The reason for the severe ASR degradation due to reverberation primarily lies in the temporal and spectral smearing effects of reverberation. Reverberation leads to a complex mixing of the neighbor sounds. Since a particular speech signal is a temporal sequence of different sound units, the impact of reverberation on a particular sound unit not only depends on the environment's RIR but also on the preceding sequence of sounds [8]. Thus, reverberation will affect different sounds differently, and even matched training will not be robust to reverberation. The objective of the proposed work is to design a set of compensation algorithms for reverberation that will sharply reduce the difference in WER between clean and the reverberant conditions.

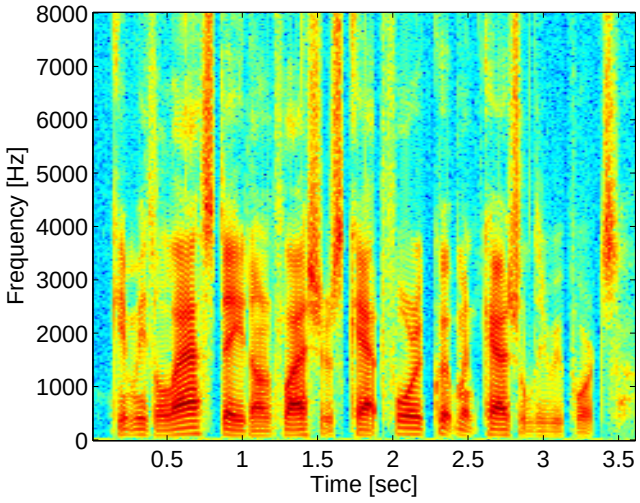
1.5 MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

In this section we briefly introduce the conventional Mel-frequency cepstral coefficients (MFCC) [14] features for an ASR system. We summarize MFCC feature extraction in Fig. 1.5. An incoming speech signal $s[n]$ is first analyzed by a filter bank which typically consists of 40 filters for 16000-Hz signal sampling frequency. The center frequencies and bandwidths of the filters are chosen to crudely mimic human ear characteristics. The $x[n, i]$ denotes the output of the i^{th} filter bank at discrete-time instant n. Next, short-time power (typically over 25 ms) is calculated for the filter-bank outputs, and $X_s[n, i]$ denotes the corresponding power coefficients for the i^{th} filter. The power coefficients $X_s[n, i]$ are fed to a logarithmic non-linearity stage to obtain the coefficients $X_l[n, i]$, which are passed through the discrete cosine transform (DCT) for dimensionality reduction to the 13-dimensional cepstral coefficients in $X_c[n, i]$, there

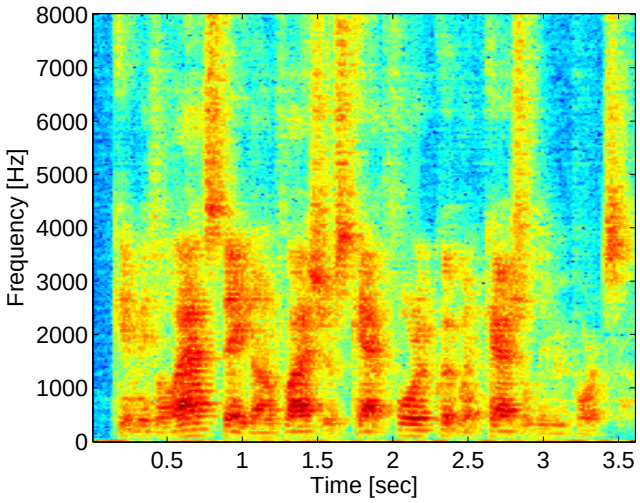
$i \in \{1 \ldots 13\}$. The X_c coefficients are also referred to as MFCC features. The ASR system operates on these MFCC features.

1.6 THESIS ORGANIZATION

The rest of this work is organized as follows. Chapter 2 provides a review of the current reverberation-compensation approaches for ASR. It also highlights the missing science in the current approaches. There, we specifically note the need to better analyze and model reverberation in the speech feature domain. We propose a new framework for analyzing and modeling reverberation in Chapter 3. We show that proposed framework is a generalization of a prior such framework. Next, we develop several algorithms using the proposed framework. Chapter 4 presents the cepstral post-filtering (CPF) algorithm for reverberation compensation, an algorithm which makes simplified assumptions about the reverberation filter to develop a post-filtering technique on the cepstral sequences. Next, we propose the maximum-likelihood-based inverse filtering (LIFE) technique in Chapter 5. There, we maximize the feature likelihood criterion to seek an inverse filter to compensate for reverberation in our proposed framework. On an appropriate ASR database, we show that the proposed approach sharply reduces the ASR WER by up to 45% relative reduction in WER for RT of 300 ms conditions. In Chapter 6 we propose a non-negative matrix factorization (NMF) approach for dereverberation. There, we use the non-negativity of speech spectra as a constraint to guide speech dereverberation. NMF provides a relative reduction of 40 to 45 percent in WER for RTs of 300 and 500 ms. In Chapter 7 we develop delta-spectral cepstral coefficients (DSCC) features for noise compensation. The DSCC features capture the dynamic spectral characteristics, and they offer a robust alternative to conventional delta-ceptral coefficients (DCC) features. DSCC features provide a 5-10 dB SNR improvement in effective SNR different additive noise conditions. In Chapter 8 we provide a framework for jointly representing reverberation and noise in the spectral domain. This framework generalizes the spectral domain reverberation framework proposed in Chapter 3. The new framework also characterizes explicitly for approximation error in the reverberation framework in Chapter 3. In Chapter 8 we also provide various combinations of the algorithms developed in this work to compensate jointly for reverberation and noise.



(a)



(b)

Fig. 1.3: *(a) Spectrogram of a clean speech signal, (b) Spectrogram of a reverberated speech sample with the RIR in Fig. 1.1(b).*

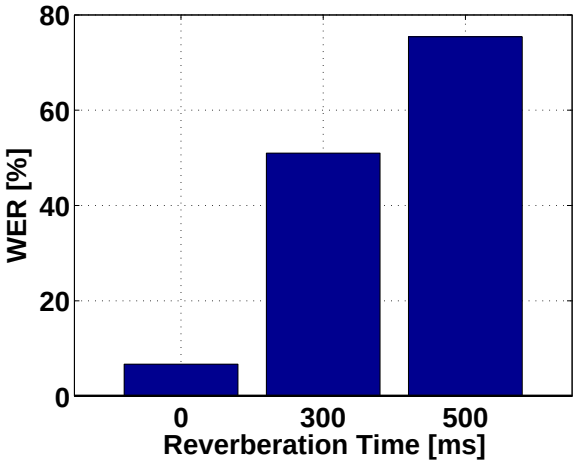


Fig. 1.4: *Baseline word error rate (WER) in reverberant conditions.*

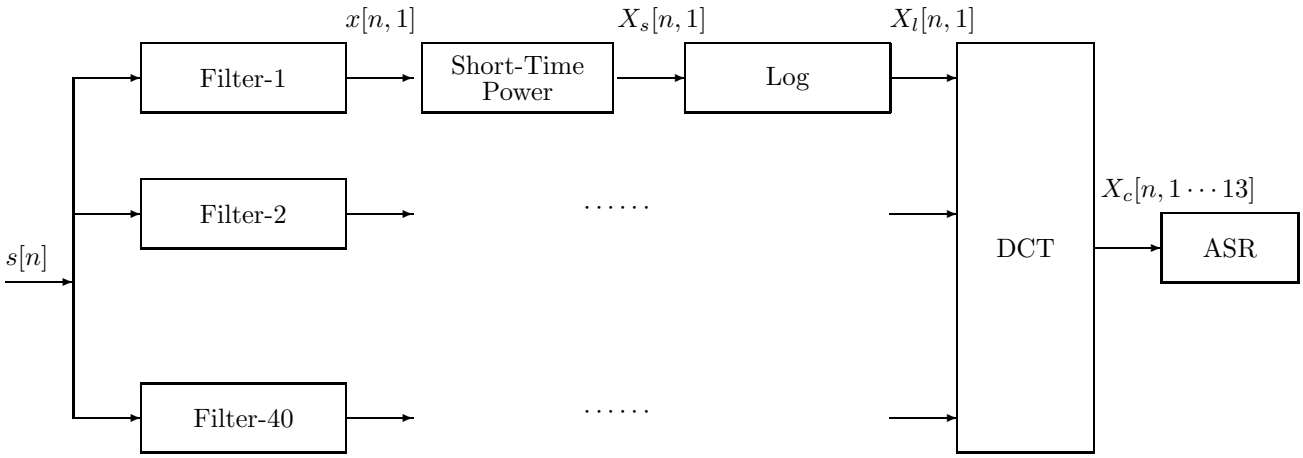


Fig. 1.5: *MFCC feature extraction for ASR.*

CHAPTER 2

LITERATURE REVIEW AND MISSING SCIENCE

In this section, we review some of the prominent dereverberation approaches for ASR. We also discuss the major limitations in the current methods and as well as missing elements in those approaches.

2.1 FEATURE-MODEL-BASED APPROACHES

The feature model-based reverberation compensation approaches parameterize the effect of reverberation on the speech features. Some of the traditional approaches model reverberation as an additive shift on the log-spectral or cepstral features, and propose algorithms to compensate for the additive shift. Next, we discuss a few specific algorithms that use this approach [15, 16].

2.1.1 Cepstral Mean Normalization

Cepstral mean normalization (CMN) is a simple and ubiquitously used algorithm for ASR. CMN was initially proposed in [17] for statistical mean normalization of speech features. Later CMN was adapted for compensating an unknown linear filtering operation whose impulse response had a very short duration (shorter than the analysis window of 25 ms for speech feature). Reverberation is also a linear filtering operation in the time domain but its impulse response extends much longer in time. Hence, the benefits of CMN do not extend to compensating for reverberation. Discussion of CMN is still important for its conceptual merit, as other algorithms such as LTLSS (discussed below) are direct variants of CMN. CMN builds on homomorphic signal analysis which transforms in the time domain to addition in the log-spectral domain. The additive components due to reverberation are relatively stationary compared to the speech components and can be compensated for by mean normalization of the

log-spectral components. CMN can work if speech feature analysis is done over long time segments but typically speech analysis is done using segments of approximately 20-30 ms. CMN modeling does not hold and provides only limited improvement in these cases.

2.1.2 Long-Term Log-Spectral Subtraction

Long-term log-spectral subtraction (LTLSS) [18] is a direct extension of CMN processing. LTLSS is essentially CMN applied to longer analysis windows on the order of 1 to 2 seconds. Even though LTLSS is first applied on longer analysis windows, speech features must be obtained from short-duration segments. This leads to the use of an analysis-synthesis framework for LTLSS. LTLSS is applied over longer analysis windows from which speech is reconstructed, and ASR features are obtained from the reconstructed speech.

2.1.3 Model based Dereverberation in Mel-Spectral Domain

A model-based dereverberation approach that utilizes the ASR acoustic models to infer the optimal speech state sequences was proposed in [19, 20]. This work models reverberation in the spectral domain and learns the reverberation parameters from data recorded in that particular environment. While CMN and LTLSS were pure feature-domain approaches, this work achieves dereverberation by finding the most likely state sequence considering both the acoustic and reverberation models. Although the algorithm works well for RIRs that are known *a priori*, it can not be generalized to unknown RIRs.

2.2 LINEAR-PREDICTION-BASED APPROACHES

A great deal of dereverberation approaches operate on the linear-prediction (LP) residual of speech signals. LP processing can be done either in the time or narrow-band frequency domains. We discuss a few approaches under the broad framework of LP processing in the following:

2.2.1 LP Residual Based Enhancement

Dereverberation has also been studied in the framework of the LP [21] residual. It was observed in [22] that the probability distribution function (PDF) of the LP residual for clean speech components is sub-Gaussian whereas the corresponding PDF for the reverberated components is approximately Gaussian. Thus, the LP residual for the reverberated segments exhibits higher entropy than that of the clean segments. Based on the above observation, they developed an algorithm to exploit the LP residual's entropy to first identify and then suppress the reverberated components. Further, the LP residual's kurtosis characteristics was utilized in [23] for dereverberation.

2.2.2 Frequency Domain Linear Prediction

Frequency domain linear prediction (FDLP) [24] also lies in the broad framework of the LP residual processing. While conventional LP residual methods operate on the time domain signals but FDLP operates in the frequency domain. Specifically, a long speech segment (1-second) is analyzed by the DCT operation. The DCT coefficients are framed by a rectangular window and LP analysis is performed on the framed DCT coefficients. The framed DCT coefficients represent a narrow-band signal for which reverberation is modeled as a constant but an unknown gain term. Reverberation compensation is achieved by gain normalizing the LP residual.

2.3 MODULATION-SPECTROGRAM-BASED APPROACHES

The modulation spectrogram of a narrow-band signal is defined as the frequency response of the signal's amplitude envelope. The modulation spectrogram for a wide-band signal, such as speech, is obtained by passing the signal through a bank of narrow-band analysis filters with the desired center frequencies, and then obtaining a frequency response for each of the narrow-band signals at filter-bank outputs. A number of reverberation-compensation approaches study the effects of reverberation on the modulation spectrogram and propose filtering schemes to compensate for reverberation.

2.3.1 Relative Spectral Processing

Relative spectral (RASTA) processing [25] is a log-spectral domain modulation-filtering scheme for dereverberation. Speech studies suggests that most of the speech information lies in the range of 4-20 Hz modulation frequencies. Based on these results, the RASTA approach makes use of an empirical filter that enhances the speech characteristics in the modulation frequencies of interest. The scheme is well motivated but its solution is at best an empirical one. In practice, RASTA filters works like CMN, in that they remove low-frequency additive components in the cepstral domain.

2.3.2 Modulation Spectrogram based Features

A new representation of speech signals in terms of its modulation [26] spectrogram was proposed in [27,28]. The work builds on the importance of the speech modulation frequencies in the range of 0-16 Hz, as was earlier noted in [25]. It was empirically observed that these modulation frequencies and the features derived from them are not only important for ASR applications but are also robust to reverberation.

2.3.3 Minimum Variance Modulation Filter

The minimum variance modulation filter (MVMF) algorithm [29] is a recent work approach to noise and reverberation compensation in the modulation frequency domain. It is motivated from and builds on the principles of RASTA processing [25]. While RASTA designed an empirical modulation filter, MVMF formulates and solves a data-driven optimization in the modulation frequency domain to minimize a distortion criterion based on modulation frequency components.

2.4 ROBUST FRONT-ENDS

A robust feature-extraction scheme is critical for ASR. While MFCCs are perhaps the most dominant features for ASR, a number of alternate feature extraction schemes [30–32] have been proposed in literature that attempt to derive speech features with inherent robustness to noise and reverberation. We discuss PNCC and AFE feature extraction schemes in the following sections.

2.4.1 Power Normalized Cepstral Coefficients

Power normalized cepstral coefficients (PNCC) [30] were recently proposed as a new feature extraction algorithm that exhibits better robustness to both noise and reverberation. The approach is motivated by principles of auditory processing. PNCC processing replaces the log nonlinearity in the MFCC features with a power-law nonlinearity that mimics some human auditory characteristics. PNCC achieves noise robustness via subtracting an estimate of the inferred distortion power. PNCC also achieves reverberation robustness via a temporal suppression stage.

2.4.2 Advanced Front-End

The ETSI advanced front-end (AFE) standard [31] is a noise robust feature extraction scheme in a client/server based distributed system framework for speech applications. The standard is primarily aimed for ASR applications on mobile phones. In the client/server model, client *i.e.* the hand-held device, performs feature processing, the resulting features are encoded and transmitted to a server which decodes the features and performs computationally heavy task of ASR decoding and sends back results to the client. This distributed framework for ASR is efficient with respect to transmission bandwidth requirement, and processing and power consumption on the client devices. The model also significantly mitigates encoding/decoding errors associated with transmitting a speech signal. Although AFE provides significant robustness to noise, it does not provide any improvement in reverberation.

2.5 INVERSE FILTERING USING A KNOWN ROOM IMPULSE RESPONSE

Some work in dereverberation has focussed on situations in which the room impulse response is known *a priori*. Even though such a scenario is not generalizable to other cases when the room impulse response is not known, deriving an inverse filter from even known room impulse response is non-trivial. This is because the room impulse response is generally non-minimum phase [33] and thus non-invertible. Further, this approach is infeasible in all cases in practise because the RIR depends on the locations of objects in the room, including sound source and receiver. If the sound source is non-

stationary then there is no unique RIR from the source to receiver, then this approach can not apply. Overall such approaches can only provide approximate solutions to the inverse problem.

2.6 TWO-STAGE PROCESSING

Recently several types of two-stage processing (*eg.* [34–37]) methods have been proposed for speech dereverberation. Since the overall room impulse response (RIR) is typically very long, to efficiently solve the problem the RIR is subdivided into early and late reflection components [38], typically marked by a 30-50 ms boundary. Two-stage methods apply individualized processing for each of the stages. The late components result in uncorrelated additive noise and are compensated for by a variant of the spectral subtraction methods. The early components are not uncorrelated, and are compensated by a variant of the inverse-filtering schemes. These methods have been found to work well in practice.

2.7 MULTIPLE-MICROPHONE-BASED APPROACHES

Nabelek and Robinson [39] compared human monaural and binaural speech perception abilities in reverberant techniques, and concluded experimentally that the binaural abilities strongly helped monaural-only perception. Consequently Speech dereverberation has a been addressed by a number of multiple-microphones-based approaches *eg.* [40–45]. Some of the work in this area has been applied to the Microsoft's Kinect gaming platform (launched in 2010). It includes an array-processing-based approach for noise and reverberation compensation. Although the focus of this thesis is single-microphone based reverberation compensation techniques, we review a few multiple-microphone-based techniques in below.

2.7.1 Interaural Time-Delay (ITD) based Techniques

A number of dereverberation approaches build on the differences in time-delay information for signals arriving at the left and right ears. For example, if the source is an anechoic room lies along is at the perpendicular bisector of the line joining the left and right ear then both the ears receive the source signal at identical instants due to

identical path lengths from the source to each of the ears. If the source is from the perpendicular bisector then the signal path lengths from the source to the two ears are different which in turn produces a different time-delay between the signals reaching the two ears [46]. This difference is referred to as interaural time delay (ITD). The techniques in [41, 42, 47] use the ITD criterion to suppress the signal components from undesired directions and consequently enhance the signal for the source. The work in [48] is another manifestation of the time-delay idea at the ears. Their work extracts zero-crossing information and makes a signal suppression decision on zero-crossings values.

2.7.2 Maximum-Likelihood Beamforming

An ASR acoustic model-based beamforming approach was proposed by Seltzer in [44] for speech dereverberation. Even though a number of beamforming approaches proposed earlier helped improve speech intelligibility in reverberant environments, they did not necessarily improve ASR accuracy. The work in [44] noted the importance of including the ASR models in the beamforming optimization, and consequently built an optimization criterion that maximizes the likelihood of the speech state sequences with respect to the acoustic models.

2.7.3 Microphone Selection Approaches

A speech frame-based score competition (FSC) in a multiple microphone setting was proposed by Jin et al. in [49, 50]. The work was targeted for speaker identification applications. At first a likelihood function is learned from clean speech frames. During testing, the likelihood scores from the different speech frames at a particular instant compete among themselves and the frame with the highest likelihood score is selected. The FSC approach does not directly compensates for reverberation but it exhibits robustness to reverberation because it can potentially select the least reverberated frame among all the reverberated frames at a particular instant.

2.8 LIMITATIONS OF CURRENT APPROACHES AND MISSING SCIENCE

In this section we discuss some of the major limitations in the current approaches and the corresponding missing elements in those approaches. We also propose research directions on the basis these missing elements in the current research.

2.8.1 Inadequate Reverberation Model

One of the most significant limitations of the current framework in reverberation compensation for ASR is the model from which the reverberation compensation algorithms operate. Many of the current approaches work on a model that represents reverberation as an additive shift in the cepstral domain and that applies variants of mean normalization techniques for reverberation compensation. These models are based on a premise that the features are obtained over long-duration segments, but typically speech features are obtained over short-duration segments (20-30 ms), for which the conventional models do not adequately represent the phenomenon of reverberation. A major contribution of our work is to provide an understanding and modeling of reverberation for the features derived from short-duration segments. We develop a new framework to parameterize the effects of reverberation. Once we have an adequate framework from representing reverberation, we can develop multiple algorithms on that framework to compensate for reverberation. A key aspect of our reverberation model is that it consists of very few modeling parameters. This significantly assists in parameter estimation and consequently in reverberation compensation because fewer parameters imply more reliable parameter estimates when only limited amount of data may be available.

2.8.2 Lack of Objective Solutions

In the development of dereverberation algorithms for ASR, there has often been a disconnect between the motivation for an algorithm and the solution proposed by the algorithm. Often the proposed solution does not adequately reflect the motivation, leading to a sub-optimal solution. Filtering schemes like RASTA were motivated by modulation characteristics but the proposed solutions are at best empirical, and in

practice RASTA processing ends up being similar to CMN. The LP-residual-based approaches may locally optimize the LP-residual characteristics of speech but may not optimize the features that go to speech recognizer. A key focus in this thesis is to design algorithms based on a defined objective function that assists ASR.

2.8.3 Lack of Theoretical Analysis and Robustness Predictions for Algorithms

The focus of many of the current approaches is more on experimental results with theoretical analysis often missing. Of course, positive experimental results are the key to the success of an algorithm but theoretical analysis provides a deeper understanding into the working of an algorithm and can also provide theoretical performance bounds on that algorithm. In our approaches we strive to provide insightful analysis into the working of our algorithms. For our dereverberation approach, we show that the algorithm is guaranteed to compensate for reverberation under certain stated conditions. In addition, very few attempts have been made in literature to predict the estimated noise-robustness capacity of algorithms. For our noise-compensation (see Ch. 7) algorithm, we provide a bound on the achievable signal-to-noise ratio (SNR) improvement from the algorithm and found a 0.93 correlation between our prediction and the observed. A key question that we raise throughout our work is the maximum benefit that an algorithm can provide under realistic experimental conditions.

2.8.4 Knowledge about the Room Environment

A number of reverberation compensation approaches require some knowledge about the room reverberation parameters. This becomes an issue with the practical deployment of the algorithm in unknown environments. A key goal of this thesis to propose completely blind reverberation compensation algorithms that do not require any knowledge about reverberation parameters. Instead of guiding our dereverberation optimization problems with prior knowledge about reverberation parameters, we propose and successfully guide our optimizations using generic speech knowledge in terms of it's feature auto-correlation sequences, feature sparsity, and feature probability distributions.

2.8.5 Feature-Domain vs. Time-Domain Processing

A great deal of past work in dereverberation for ASR has been directly borrowed from the dereverberation work in the speech enhancement community. The focus in the speech enhancement work is to improve speech intelligibility for consumption by humans and in practice these methods provide limited improvement for ASR. ASR is essentially a feature-based pattern matching algorithm where the features are derived from nonlinear processing on the time domain signals. The nonlinearity implies that the optimal criterion in the time domain may not be optimal in the feature domain. In our work, we build feature-domain optimization criterions such that ASR can directly benefit due to better feature matching.

2.8.6 Unified Model for Environments with both Noise and Reverberation

In our work we provide a new perspective on the problem of a joint noise and reverberation modeling and experimentally verify its benefits on speech utterances that are affected by both noise and reverberation. Many of the past such joint models were derived from premises of noise, as against reverberation, being the most dominant source of degradation. That worked well for noise-only conditions but did not adequately represent reverberation and correspondingly it neither worked for reverberation-only nor for both noise-and-reverberation conditions. We instead derive our unified model from our reverberation-only model that adequately represents reverberation and then we naturally generalize it to a unified model for both noise and reverberation conditions. The unified model can also be used to encapsulate the additive error in our reverberation model, and highlights that a better dereverberation technique should include denoising as an intermediate step.

CHAPTER 3

MODELING REVERBERATION IN THE FEATURE DOMAIN

This chapter is devoted to modeling and representing reverberation for the purpose of automatic speech recognition (ASR). One of the most significant limitations of the current reverberation-compensation framework for ASR is the model on which the reverberation-compensation algorithms operate. Since the reverberation-compensation algorithms are derived from a model, the compensation algorithms will be sub-optimal if the model itself is inadequate. In this chapter we first present a conventional model of reverberation in the cepstral feature domain. We highlight its limitations and present a new framework for representing reverberation in the feature domain. The proposed framework will be used in subsequent chapters for designing reverberation compensation algorithms.

3.1 CONVENTIONAL FEATURE-DOMAIN REVERBERATION MODEL

In this section we present a conventionally-used model of reverberation with respect to the MFCC features in Fig. 1.5. This model builds on the physical model of reverberation in Fig. 1.2 and the background discussion in Sec. 1.5 in which we presented the conventional MFCC features for ASR. The time-domain reverberation model in Fig. 1.2 represented reverberation as an LTI system in terms of a filter $h[n]$. MFCC features in Fig. 1.5 were obtained by analyzing the speech signal $s[n]$ in terms of a filter bank. Fig. 3.1 shows the same analysis of the input signal $s[n]$ in presence of reverberation. There, $s[n]$ will first undergo a linear filtering operation by $h[n]$, the corresponding output $\tilde{s}[n]$ will then be analyzed by the filter bank. The variable $y[n, i]$ refers to the output of the i^{th}-filter. Both $h[n]$ and the i^{th}-filter constitute linear filtering operations, and since linear filtering is commutative, the order of the filters Filter-i and $h[n]$ can be

interchanged without affecting the channel output $y[n,i]$. We illustrate an interchange of the above filters in Fig. 3.2. Referring to Fig. 3.2, $x[n,i]$ is the i^{th}-filter output

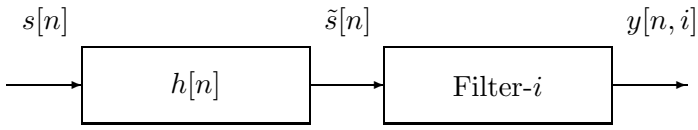


Fig. 3.1: *A physical model of reverberation.*

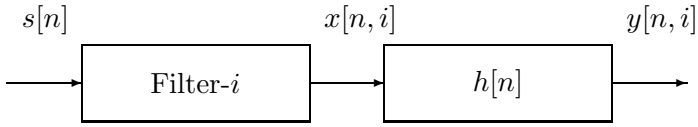


Fig. 3.2: *An equivalent representation for $y[n,i]$ in Fig. 3.1.*

in the absence of reverberation and $y[n,i]$ is the corresponding output in presence of reverberation. Thus we have a convolution operation in $y[n,i] = x[n,i] * h[n]$. Note that the convolution is performed over the discrete-time index n. Speech features are derived from the signal spectral contents and since convolution in time is multiplication in frequency, we obtain

$$\mathcal{Y}[k,i] = \mathcal{X}[k,i] \cdot \mathcal{H}[k] \tag{3.1}$$

where k is an index for frequency and the symbols $\mathcal{Y}[k,i]$, $\mathcal{X}[k,i]$,and, $\mathcal{H}[k]$ respectively denote the discrete fourier transform (DFT) of $y[n,i]$, $x[n,i]$, and, $h[n]$. We reiterate that the signal $x[n,i]$ and the coefficients $\mathcal{X}[k,i]$ correspond to clean speech $s[n]$ and that $y[n,i]$ and $\mathcal{Y}[k,i]$ correspond to reverberated speech $\tilde{s}[n]$. Note that except for $\mathcal{H}[k]$, the DFT terms include the index i corresponding to the i^{th}-filter. Currently, it is important to note that the DFT analysis is performed over the entire duration of the signals. This is in contrast to the feature analysis in Fig. 1.5, where feature analysis is performed over short segments ($\sim$25 ms) by evaluating a short-term power in that segment. Next, noting that currently we operate on the entire signal, we evaluate

signal power at the output of the i^{th}-filter in:

$$P_y[i] = \frac{1}{N_y}\sum_n y[n,i]^2 = \frac{1}{N_y^2}\sum_k |\mathcal{Y}[k,i]|^2 \quad (3.2)$$

$$= \frac{1}{N_y^2}\sum_k |\mathcal{X}[k,i]|^2 \cdot |\mathcal{H}[k]|^2 \quad (3.3)$$

$$\approx \frac{1}{N_y^2}\sum_k |\mathcal{X}[k,i]|^2 \cdot \hat{P}_h[i] \quad (3.4)$$

$$= P_x[i] \cdot \hat{P}_h[i] \quad (3.5)$$

where, $P_x[n,i]$ and $P_y[n,i]$ respectively denote the signal power corresponding to $x[n,i]$ and $y[n,i]$. Note that the power evaluation is done over the entire duration of $y[n,i]$ whose length is N_y. In deriving (3.5), we first applied Parseval's theorem in (3.2), incorporated (3.1) in (3.3). The approximation in (3.4) can be justified if the signal $x[n,i]$ exhibits a narrow bandwidth as it is the output of the i^{th}-filter, which is a band-pass filter with a narrow bandwidth. Thus, for each i, most of the k terms in $\mathcal{X}[k,i]$ will be approximately 0. Hence, the summation over k in (3.3) will run only over a small connected range of k. Finally, we derive (3.4) by assuming that the frequency response of $h[n]$ is approximately constant over a narrow frequency range of $x[n,i]$. Equation (3.5) represents the overall power of $y[n,i]$ in terms of a product operation over the corresponding power of $x[n,i]$.

Next, the feature extraction procedure applies a logarithmic operation on the signal power content, so we obtain

$$\log(P_y[i]) = \log(P_x[i]) + \log(\hat{P}_h[i]) \quad (3.6)$$

which in the log-spectral domain, represents the effects of reverberation as an additive shift operation. The DCT operation is next applied on the log-spectral components to derive cepstral features. Since the DCT is a linear operation, the additive components in the log-spectral domain are still additive in the cepstral domain. Thus, the conventional cepstral-domain model represents reverberation as an additive shift in the cepstral domain. We reiterate that the model has been derived for feature analysis performed on the entire duration of the signal. Note that in Fig. 1.5, the cepstral features are evaluated over short signal segments so the model needs to be extended for features derived from short segments. The conventional model simply assumes

that the additive shift model derived for the entire duration of the signal holds for the features derived from short segments as well. Finally the conventional approach relates the cepstral features for the signals $y[n,i]$ and $x[n,i]$ in below:

$$Y_c[n,j] = X_c[n,j] + H_c[j] \tag{3.7}$$

where, $Y_c[n,j]$ and $X_c[n,j]$ are respectively the j^{th} cepstral feature for the signals $\tilde{s}[n]$ (reverberated speech) and $s[n]$ (clean speech), and $H_c[j]$ is a constant representing the effect of reverberation. We show the above reverberation model in Fig. 3.3. Thus

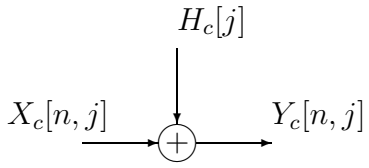


Fig. 3.3: *Conventional reverberation model for the j^{th} cepstral feature.*

the conventional model represents reverberation as an unknown but constant additive shift in the cepstral domain. This is the model of reverberation used in many studies, where reverberation compensation is done by a variant of cepstral mean normalization (CMN), which removes the additive bias due to reverberation. Although CMN provides significant robustness to ASR and is ubiquitously used in all ASR systems, the ASR accuracy with CMN processing is still limited. The key limitation in CMN is due to limitations in the model in Fig. 3.3. The model holds only if the feature analysis is performed over the entire duration of the signal. The model is approximately valid if the the duration of the filter $h[n]$ is significantly small in comparison to the feature analysis duration ($\sim$25 ms) and has been found to be useful in mitigating the impulse responses associated with speech recording devices like microphones. Since the RT of a room can last hundreds of milliseconds, the conventional cepstral domain model does not accurately hold for representing the effects of reverberation. In the next few sections we provide a new framework for modeling reverberation in the cepstral domain and show it to be a correction over the conventional model in Fig. 3.3.

3.2 MATHEMATICAL MODEL OF REVERBERATION IN THE SPECTRAL DOMAIN

In this section we provide a new model for representing reverberation in the spectral domain, and later in Sec. 3.3.1 we extend the model to cepstral domain. We build our model from the development so far in Fig. 3.2. We incorporate the model in Fig. 3.2 in the ASR system from Fig. 1.5, resulting in Fig. 3.4. We note that since our model will be derived from the signal flow in Fig. 3.4, it will incorporate the issue of speech features analysis being performed over short signal segments. This is in contrast to the assumption of feature analysis being performed over the entire duration of the signal in the conventional model in Sec. 3.1. The index i in $\{Y_s[n,i], Y_l[n,i]\}$ represents the

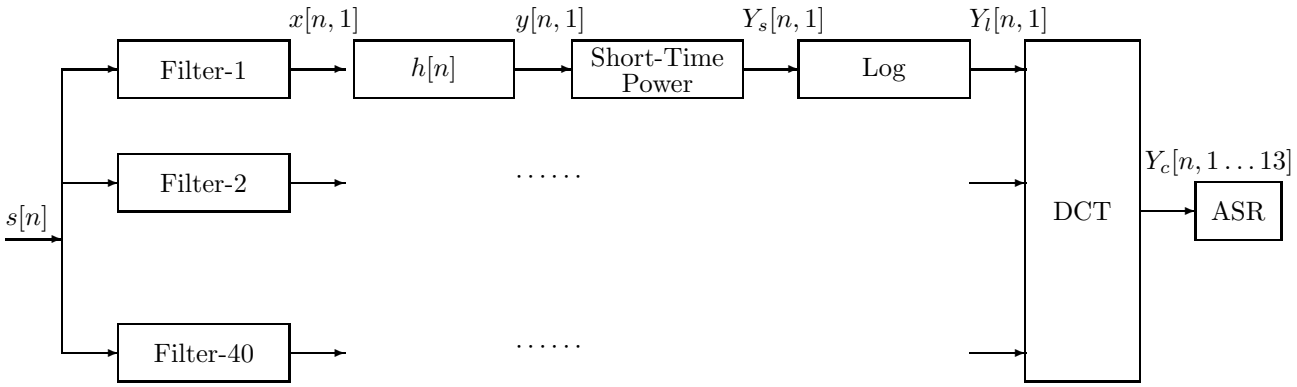


Fig. 3.4: *Incorporating Reverberation in MFCC feature extraction.*

outputs corresponding to the i^{th}-filter in the filter bank. $Y_c[n,1\ldots13]$ represents the 13-dimensional cepstral features over the discrete-time index n. Since signal spectral analysis is performed individually for each of the filters in the filter bank, we discuss further modeling and analysis in terms of only one of the filter, specifically the Filter-1. The analysis for the first filter can easily be extended to the rest of the filters in the filter bank. Next, for brevity of notation we use the symbols $\{x[n], y[n], Y_s[n], Y_l[n]\}$ to represent the single-dimensional sequences for the first filter in the filter bank. From Fig. 3.4, we note that in presence of reverberation the ASR system will operate on the cepstral coefficients Y_c that are derived from spectral coefficients Y_s. The objective of this section is to relate Y_s to the corresponding clean spectral coefficients, *i.e.*, X_s in Fig. 1.5 even when the signal analysis is performed over short segments. We begin

from the linear-filtering model relating $x[n]$ and $y[n]$ and first relate the $x^2[n]$ and $y^2[n]$ terms. The short-time power for $x[n]$ and $y[n]$ will be obtained by averaging $x^2[n]$ and $y^2[n]$ over a short duration. We begin from $y[n] = \sum_{l=0}^{N_h-1} h[l]\, x[n-l]$ to show that

$$\begin{aligned} y^2[n] &= \sum_l \sum_m h[l]h[m]x[n-l]x[n-m] \\ &= \sum_l h^2[l]\, x^2[n-l] + \underbrace{\sum_l \sum_{m \neq l} h[l]h[m]x[n-l]x[n-m]}_{E[n]} \end{aligned} \tag{3.8}$$

Thus, $y^2[n]$ is composed of two broad components. One of the components is a linear filtering operation over $x^2[n]$ and the other is a noise-like additive component, $E[n]$. The $E[n]$ term represents error in the linear filtering approximation of $y^2[n]$. Next, we evaluate short-duration power by averaging over a short duration of the $y^2[n]$ terms. Since the speech feature analysis is typically performed over $\sim$25 ms, the short duration power for a particular n will be evaluated by weighted-average over the past 25 ms of $y^2[n]$. The power averaging may be done by uniformly weighting all the $y^2[n]$ terms. The uniform weighting corresponds to using a rectangular window in the feature analysis. The power averaging may also be done for any other window shape, and is most commonly performed using a hamming window. We describe the weighted power averaging operation in terms of an $\mathbb{A}[.]$ operator. The $\mathbb{A}[.]$ operator is a linear operator which for time-instant n, windows the past 25 ms $y^2[n]$ coefficients according to a specified window function and calculates an average over the windowed $y^2[n]$ terms. Using the linear $\mathbb{A}[.]$ operator over a short segment of $y^2[n]$, we have:

$$\begin{aligned} \mathbb{A}[y^2[n]] &= \mathbb{A}[\sum_l h^2[l]\, x^2[n-l] + E[n]] \\ &= \sum_l h^2[l]\, \mathbb{A}[x^2[n-l]] + \mathbb{A}[E[n]] \end{aligned} \tag{3.9}$$

where we invoked the linearity of $\mathbb{A}[.]$. We formalize our spectral-domain model in

$$\begin{aligned} Y_s[n] &= \underbrace{X_s[n] * H_s[n]}_{\hat{Y}_s[n]} + E_s[n] \\ \text{where, } Y_s[n] &= \mathbb{A}[y^2[n]],\ X_s[n] = \mathbb{A}[x^2[n]],\ H_s[n] = h^2[n],\ E_s[n] = \mathbb{A}[E[n]] \\ &\text{and for small } E_s[n],\ Y_s[n] \approx \hat{Y}_s[n] \end{aligned} \tag{3.10}$$

Eq. (3.10) represents $Y_s[n]$ in terms of a convolution over $X_s[n]$ and an additive $E_s[n]$ term, where the index n represented discrete-time sample instants. For ASR applications, $Y_s[n]$ is invariably downsampled. In (3.10) we derived a spectral-domain model before the downsampling stage but we assume it to hold in the downsampled domain as well. Thus, in either the pre-downsampling or post-downsampling stage, we approximate a reverberated spectral feature sequence as a convolution over the corresponding clean feature sequence. Further, we define the following distortion criterion to quantify the model approximation in $Y_s[n] \approx \hat{Y}_s[n]$:

$$D_s = 10 \log_{10} \left(\frac{\mathbb{P}[X_s[n]]}{\mathbb{P}[E_s[n]]} \right) \tag{3.11}$$

where, $\mathbb{P}[.]$ is the conventional power operator for a discrete-time sequence. A high value for D_s indicates smaller approximation error and thus a better fit of the model. We provide an empirical demonstration of the spectral domain model in (3.10) in Fig. 3.5 and Fig. 3.6. There, we applied a hamming window with duration 25 ms and window shift of 10 ms to demonstrate the model distortion ratio D_s in post-downsampled domain. We plot D_s (without $10 \log_{10}$) for each of the 40 Mel-channels in Fig. 3.5. In Fig. 3.6, we plot D_s in $10 \log_{10}$ domain (as in (3.11)). In both of the above plots, we plot the distortion ratio for a power-spectral sequence as well as for the corresponding magnitude-spectral sequence. We see that the D_s ratio is larger in the magnitude-spectral domain than in power-spectral domain, specifically the distortion is about 2-3 times higher in the magnitude-spectral domain than in power-spectral domain. Thus, the magnitude-spectral model offers a better approximation for $Y_s[n] \approx \hat{Y}_s[n]$.

In Fig. 3.7, we provide a demonstration of the model in (3.10) with respect to its fit at 4 different Mel channels. Note that we already noted the corresponding distortion levels in Fig. 3.5.

In this section we provided a model of reverberation in the spectral domain. The model was derived according to the conventional feature extraction steps in Fig. 3.4. We note that the model derivation incorporates the fact that the feature analysis is performed over short signal segments. In the next section we extend the model to the log spectral domain and also to the cepstral domain.

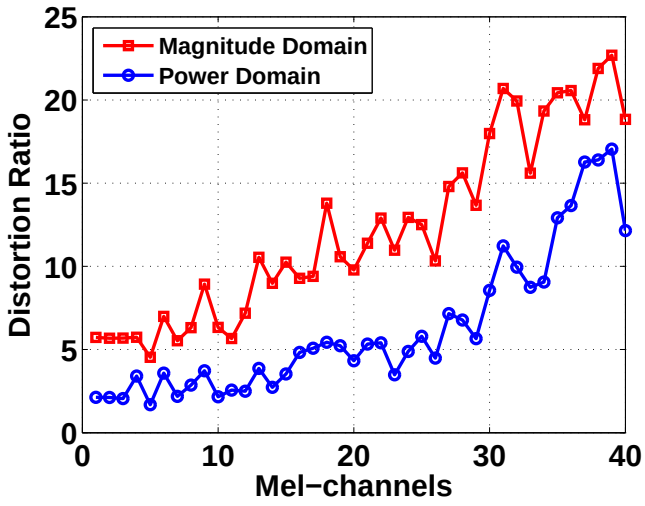


Fig. 3.5: *An empirical evaluation of the distortion ratio in* (3.11) *for each of the Mel-channels.*

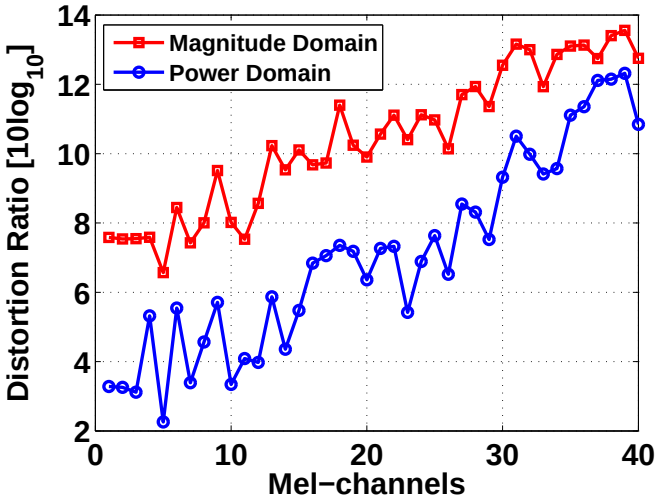


Fig. 3.6: *An equivalent plot of the distortion ration in Fig. 3.5 in* $10\log_{10}$ *domain.*

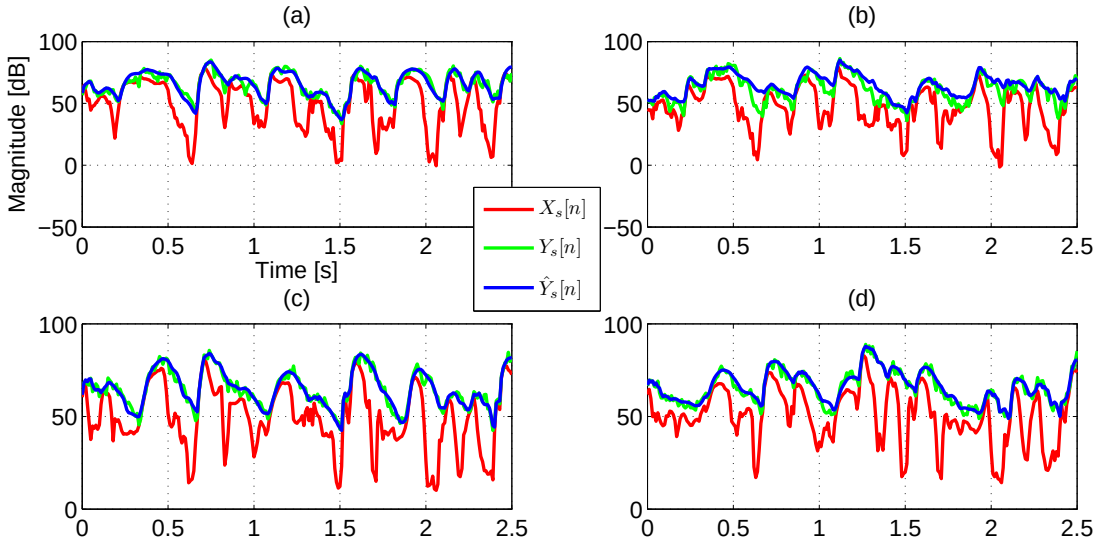


Fig. 3.7: *A demonstration of the approximation error in the spectral-domain model in* (3.10) *for (a)* 7^{th} *Mel-channel centered at 508 Hz, (b)* 14^{th} *Mel-channel centered at 1060 Hz, (c)* 21^{st} *Mel-channel centered at 1860 Hz, (d)* 28^{th} *Mel-channel centered at 3030 Hz.*

3.3 MATHEMATICAL MODEL OF REVERBERATION IN THE LOG-SPECTRAL AND CEPSTRAL DOMAINS

In this section we build on the spectral-domain reverberation model in (3.10) and derive a reverberation model in the log-spectral domain. We begin from the linear filtering model in (3.10) to obtain $Y_s[n] = \sum_m H_s[m]X_s[n-m]$. For subsequent use, we define the sum of filter-taps as S_Δ and the normalized filter-taps as $\acute{H}_s$

$$S_\Delta = \sum_m H_s[m], \qquad \acute{H}_s[m] \triangleq H_s[m]/S_\Delta$$

Using the above notations we obtain an equivalent representation of Y_s in

$$Y_s[n] = S_\Delta \sum_m \acute{H}_s[m]X_s[n-m]$$

We take the log operation on Y_s to obtain log-spectral components:

$$\underbrace{\log(Y_s[n])}_{Y_l[n]} \geq \underbrace{\log(S_\Delta)}_{L_\Delta} + \sum_m \acute{H}_s[m] \underbrace{\log(X_s[n-m])}_{X_l[n-m]}$$

there we applied Jensen's inequality and for brevity we defined $\{Y_l[n], X_l[n], L_\Delta\}$ as highlighted above. Note that Jensen's inequality requires that $\sum_m \acute{H}_s[m] = 1$ and $\acute{H}_s[m] \geq 0$. Finally we have

$$Y_l[n] = \underbrace{L_\Delta + \sum_m \acute{H}_s[m] X_l[n-m]}_{\hat{Y_l}[n]} + E_l[n]$$

where, $E_l[n]$ incorporates the approximate error due to the Jensen's inequality. If the error term $E_l[n]$ is small, we can obtain:

$$Y_l[n] \approx \hat{Y_l}[n] = L_\Delta + \sum_m \acute{H}_s[m] X_l[n-m] \tag{3.12}$$

Our reverberation model (3.12) in the log-spectral domain models reverberation in terms of two components, a liner-filtering operation and then an additive shift.

We empirically study the fit of the log-spectral domain model in (3.3) with respect to a distortion ratio criterion in the log-spectral domain. The aforesaid distortion criterion is defined in parallel to the spectral domain distortion criterion in (3.11) as shown below:

$$D_l = 10 \log_{10} \left(\frac{\mathbb{P}[X_l[n]]}{\mathbb{P}[E_l[n]]} \right) \tag{3.13}$$

In parallel to Fig. 3.6, Fig. 3.8 and Fig. 3.9 demonstrate a fit of the log-spectral model in downsampled log-spectral domain. We see that the model achieves a distortion level in the range of 16-19 dB. Next, we extend the reverberation model to the cepstral domain.

3.3.1 Mathematical model of Reverberation in the Cepstral Domain

The MFCC features fed to an ASR system are derived by DCT operation on the log-spectral features. In this section we extend the derived log-spectral domain in (3.12) to the cepstral domain. The DCT is a linear operation, so a reverberation model in

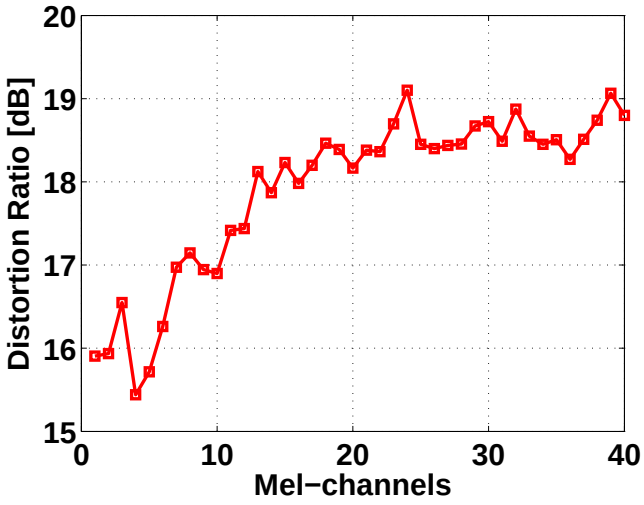


Fig. 3.8: *An empirical evaluation of the distortion ratio in* (3.13) *for each of the Mel-channels.*

the log-spectral domain directly extends to the cepstral domain. Thus in the cepstral domain we obtain

$$Y_c[n] = C_\Delta + \sum_m H_c[n] X_c[n-m] \tag{3.14}$$

where Y_c and X_c are respectively the cepstral sequences for y and x. We represent our cepstral domain models in Fig. 3.10. There, Fig. 3.10(a) represents a generic cepstral domain reverberation model and Fig. 3.10(b) represents a ceptral domain model for j^{th} cepstral feature.

Fig. 3.10 represents our new framework to study reverberation in the cepstral domain representation of the speech features. The model represents reverberation effects in the cepstral domain in terms of two components. One of the components is a linear filtering operation and the other is a constant additive component. This model can be compared with the conventional reverberation model in Fig. 3.3. The current model can be seen as an extension of the conventional model as it extends the conventional model to include a linear filtering operation in the model. It is also worthwhile comparing our reverberation model in the cepstral domain with that of the joint channel and noise model in [51–53]. It is very interesting to note that despite the vast differences in the domain and the scope of two models, they consist

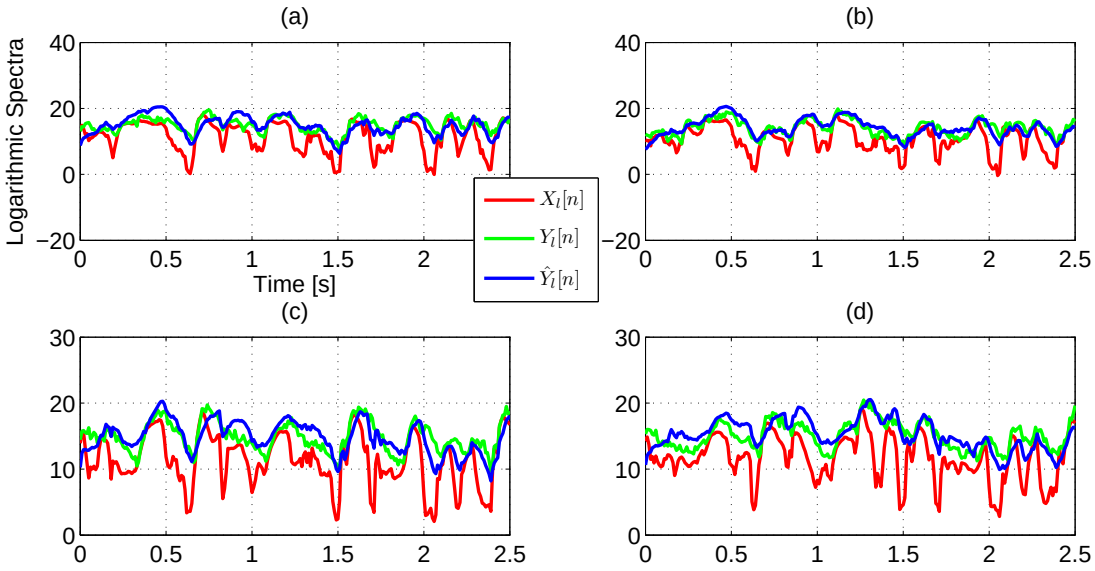


Fig. 3.9: *A demonstration of the approximation error in the log-spectral domain model in* (3.12) *for (a)* 7^{th} *Mel-channel centered at 508 Hz, (b)* 14^{th} *Mel-channel centered at 1060 Hz, (c)* 21^{st} *Mel-channel centered at 1860 Hz, (d)* 28^{th} *Mel-channel centered at 3030 Hz.*

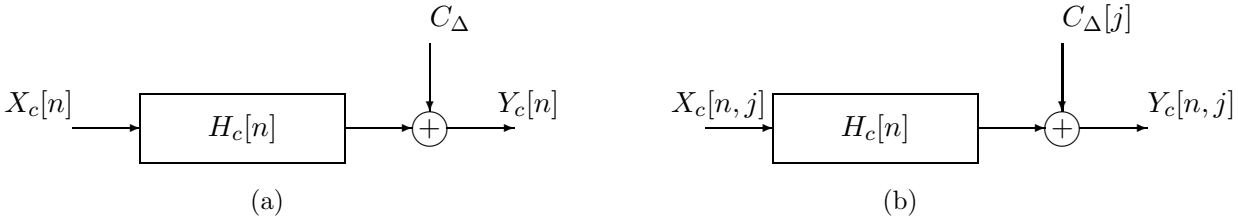


Fig. 3.10: *(a) A generic reverberation model in the cepstral Domain, (b) Reverberation model for the* j^{th} *cepstral feature.*

of identical fundamental blocks. Finally, we note that the overall reverberation model is a nonlinear model. The nonlinearity stems due to the constant additive term which is a nonlinear operation.

3.4 DISCUSSION

In this chapter we first presented a conventional model of reverberation. We noted that the model is strictly valid only if the speech analysis is performed over the entire duration of the signal. The model is approximately valid if the duration of the impulse response is small compared to the feature-analysis duration. Typically speech analysis is performed over short segments (25 ms) for which the conventional model does not adequately represent reverberation that extends up to hundreds of milliseconds. Any compensation approach based on the inadequate model will also be sub-optimal. An accurate understanding and modeling of reverberation for speech features is therefore critical to the success of an algorithm and the key focus of this chapter was to develop an adequate model of reverberation. In our work we derived reverberation representation at the different stages in feature extraction, *i.e.*, in the spectral and cepstral domains. Our model incorporates the fact that feature analysis is performed over short segments. The subsequent chapters in this thesis will use the framework proposed in this chapter to derive reverberation compensation algorithms.

CHAPTER 4

CEPSTRAL POST-FILTERING

In Chapter 3 we provided a model for representing reverberation in the cepstral feature domain. In this chapter we will build on the proposed model and provide an algorithm for compensating reverberation in that domain. The proposed algorithm, called cepstral post filtering (CPF) [1], is shown in Fig. 4.1. There, X_c and Y_c respectively indicate the clean and reverberated cepstral features and H_c models the room reverberation in terms of a linear filter. Thus, the ASR training will be done on the X_c features, while ASR testing will operate on the Y_c features. This leads to a mismatch between the training and testing features and correspondingly a loss in ASR performance.

4.1 CPF FORMULATION

In cepstral post-filtering work we characterize the mismatch in terms of a mean-squared criterion and attempt to identify a cepstral filter that minimizes the mismatch between the clean and reverberated features. Specifically, in Fig. 4.1 we design a post-filtering scheme to minimize mismatch between the cepstral features X_c and Y_c. Thus, our objective is to design an FIR filter $P[n]$ to minimize $e[n]$ in the mean-squared error sense. It can be seen that

$$\begin{aligned} e[n] &= W_c[n] - Z_c[n] \\ &= P[n] * X_c[n] - P[n] * H_c[n] * X_c[n] \end{aligned} \tag{4.1}$$

The length of filter $P[n]$ is assumed to be N_p. We seek to obtain the optimal filter $P[n]$ which minimizes the mean square error of $e[n]$.

[1] This work was published in [54].

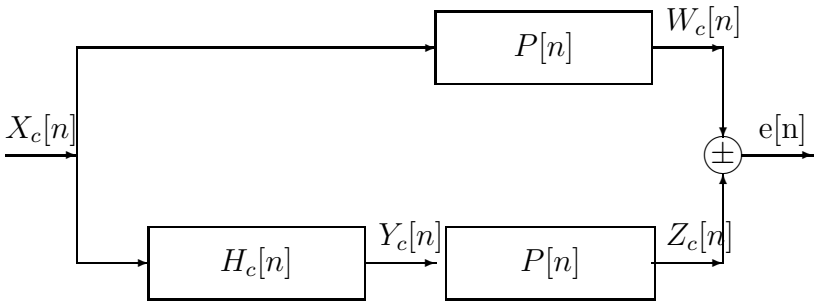


Fig. 4.1: Cepstral post-filtering for reverberation compensation.

4.2 CPF OPTIMIZATION

In this section we derive the optimal filter $P[n]$ as noted in Fig. 4.1. For convenience we rewrite the effect of $H_c[n]$ as

$$\begin{aligned} W_c[n] - Z_c[n] &= P[n] * \big(X_c[n] - H_c[n] * X_c[n]\big) \\ &= P[n] * \big(X_c[n] - H_c[0]X_c[n] - \underbrace{\sum_{i=1}^{N_h-1} H_c[i]X_c[n-i]\big)}_{r[n]} \\ &= (1 - H_c[0])\, P[n] * X_c[n] - P[n] * \big(\sum_{i=1}^{N_h-1} H_c[i]X_c[n-i]\big) \\ &= \underbrace{(1 - H_c[0])\, W_c[n]}_{f[n]} - \underbrace{P[n] * \big(\sum_{i=1}^{N_h-1} H_c[i]X_c[n-i]\big)}_{d[n]} \end{aligned} \tag{4.2}$$

where N_h represents the length of $H_c[n]$. The effect of the filter $\mathcal{H}_c(z)$ (z-transform of $H_c[n]$) is decomposed into the following two components to represent the direct and the reflected signal components, (a) the coefficient $H_c[0]$ encapsulates the direct signal component, (b) the filter coefficients $\{H_c[n]\}$ $\forall n \geq 1$ account for the reflected and attenuated signal components. We note that the filter coefficients have been time-delayed to account for any delay in the direct signal component. The unknown delay may also be incorporated in $X_c[n]$ itself. Thus, in the absence of reverberation, the filter $\mathcal{H}_c(z)$ reduces to a multiplication by $H_c[0]$ and in the presence of reverberation,

the filter $\mathcal{H}_c(z)$ introduces additional delayed and attenuated components in $r[n]$. In (4.2), we also introduced $f[n]$ and $d[n]$ for later use. The difference between $W_c[n]$ and $Z_c[n]$ is due to two terms, $f[n]$ and $d[n]$. The $f[n]$ term is simply a scaled version of $W_c[n]$, which can be normalized by scale normalization. The actual error between $W_c[n]$ and $Z_c[n]$ is due to the $d[n]$ term and our next objective is to compensate for that error. We further simplify $d[n]$ as below:

$$\begin{aligned} d[n] &= P[n] * \big(\sum_{i=1}^{N_h-1} H_c[i]X_c[n-i]\big) \\ &= P[n] * \big(\sum_{j=0}^{N_h-2} H_c[j+1]X_c[n-1-j]\big), \text{ substituting } j = i-1 \text{ or } i = j+1 \\ &= P[n] * \underbrace{\big(\sum_{j=0}^{N_h-2} G_c[j]X_c[n-1-j]\big)}_{d_u[n]} \end{aligned} \tag{4.3}$$

where, for convenience we defined a new filter $G_c[n]$ where the filter-taps of $G_c[n] = H_c[n+1] \ \forall \ n \in \{0 \dots (N_h-2)\}$. Thus filter $G_c[n]$ is a (N_h-1) length FIR filter and is essentially the filter $H_c[n]$ whose filter-taps have been advanced by a single tap and the filter-tap $H_c[0]$ set to 0. Our objective now is to minimize the error $d[n]$ in the minimum squared error sense. Thus, we seek the optimal filter $P[n]$ which, when applied to both X_c and Y_c, minimizes the mean square compensated distortion $d[n]$ as defined above. So we minimize the expected distortion, $\mathbb{E}[d^2[n]]$. Using (4.3), we obtain $\mathbb{E}[d^2[n]]$ as below:

$$\mathbb{E}[d^2[n]] = \sum_{0 \le i,j \le N_p-1} P[i]P[j]\mathbb{E}[d_u[i]d_u[j]] \tag{4.4}$$

where, N_p represented number of filter taps in the filter $P[n]$.The terms $\mathbb{E}[d_u[m]d_u[n]]$ in above can be obtained by using (4.3) as below:

$$\begin{aligned} \mathbb{E}[d_u[m]d_u[n]] &= \sum_{0 \le i,j \le N_h-2} \mathbb{E}[G_c[i]G_c[j]X_c[m-1-i]X_c[n-1-j]] \\ &\cong \sum_{0 \le i,j \le N_h-2} \mathbb{E}[G_c[i]G_c[j]X_c[m-i]X_c[n-j]] \end{aligned} \tag{4.5}$$

where the equivalence in the equation above is due to the expectation operation. Such an optimization will in general be a function of the filter $G_c[n]$ which is of course completely unknown. Thus, to proceed further we need to solve the problem under certain reasonable assumptions. Even though, $G_c[n]$ is completely unknown, we can make certain assumptions about the frequency response characteristics of the filter $G_c[n]$. Since $G_c[n]$ operates on the individual cepstral features X_c which possess a narrow bandwidth, we assume that the filter $G_c[n]$ has a flat gain characteristic for the narrow-band features X_c. We specifically assume that the deterministic autocorrelation sequence (ϕ_{GG}) of the filter $G_c[n]$ to be $\phi_{GG}[n] = 0,\ \forall\ n \neq 0$. Thus, we assume that

$$\mathbb{E}[G_c[i]G_c[j]] = \sigma^2\delta[i-j], \quad \sigma^2 \neq 0 \tag{4.6}$$

with δ being Kronecker delta, we can obtain $\mathbb{E}[d_u[m]d_u[n]]$ in (4.5) as

$$\mathbb{E}[d_u[m]d_u[n]] = (N_h - 1)\sigma^2 R_X[n-m] \tag{4.7}$$

where R_X is the autocorrelation sequence of X_c. Substituting (4.7) into (4.4), we obtain

$$\mathbb{E}[d^2[n]] = (N_h - 1)\sigma^2 \sum_{0 \leq i,j \leq N_p - 1} P[i]P[j]R_X[i-j] \tag{4.8}$$

We can differentiate (4.8) with respect to $P[n]$ to find the optimal $P[n]$ but this will result in the optimal P being $\mathbf{0}$: if all the elements in $P[n]$ are equal to 0, all features in X_c and Y_c will be mapped to 0, and the mean square distortion $\mathbb{E}[d_u[n]^2]$ will always be zero as well. While this is clearly the optimal solution in the mathematical sense, it is not a useful solution. In order to avoid the degenerate solution $P = \mathbf{0}$ we further constrain filter $P[n]$:

$$\sum_{j=0}^{N_p-1} P[j] = 1 \tag{4.9}$$

Since the filter $P[n]$ operates on energy-based coefficients, (4.9) imposes an energy constraint and normalizes the filter energy.

To minimize $\mathbb{E}[d_c^2[n]]$ in (4.8) under (4.9), we construct a Lagrangian optimization criterion as below:

$$\Lambda(\boldsymbol{p}, \lambda) = (N_h - 1)\sigma^2 \sum_{0 \leq i,j \leq N_p - 1} P[i]P[j]R_X[i-j] + \lambda(\sum_{j=0}^{N_p-1} P[j] - 1) \tag{4.10}$$

Differentiating (4.10) with respect to $[P, \lambda]$ and equating the differentials to zero, we can obtain the optimal filter $P[n]$ as below:

$$\begin{bmatrix} R_X[0] & R_X[1] & \dots & R_X[N_p-1] & 1 \\ R_X[1] & R_X[0] & \dots & R_X[N_p-2] & 1 \\ \dots & \dots & \dots & \dots & \dots \\ R_X[N_p-1] & R_X[N_p-2] & \dots & R_X[0] & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} P[0] \\ P[1] \\ \dots \\ P[N_p-1] \\ \lambda' \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix} \tag{4.11}$$

The filter taps of the optimal filter $P[n]$ can be compactly obtained in:

$$P = \frac{R_{XX}^{-1}\mathbf{1}}{\mathbf{1}^T R_{XX}^{-1}\mathbf{1}} \tag{4.12}$$

where R_{XX} is the auto-correlation matrix of X_c. The obtained filter $P[n]$ is applied across all reverberation conditions, The obtained filter is therefore invariant to the reverberation condition.

Finally using the filter $P[n]$ from (4.12) in (4.2), we obtain approximately:

$$\begin{aligned} W_c[n] - Z_c[n] &\approx (1 - H_c[0])W_c[n] \\ Z_c[n] &\approx H_c[0]W_c[n] \end{aligned} \tag{4.13}$$

Thus assuming that the term $d[n]$ in (4.2) is compensated for by the filter $P[n]$, $Z_c[n]$ is approximately a scaled version of $W_c[n]$. Note that $H_c[0]$ is also completely unknown and thus we account for the unknown scale factor by scale normalization using the established cepstral variance normalization procedures.

4.2.1 CPF Assumption

The CPF processing derived in Sec. 4.2 is built upon the assumption in (4.6). In this section we specifically discuss and understand the validity of that assumption. In CPF

processing we assume that:

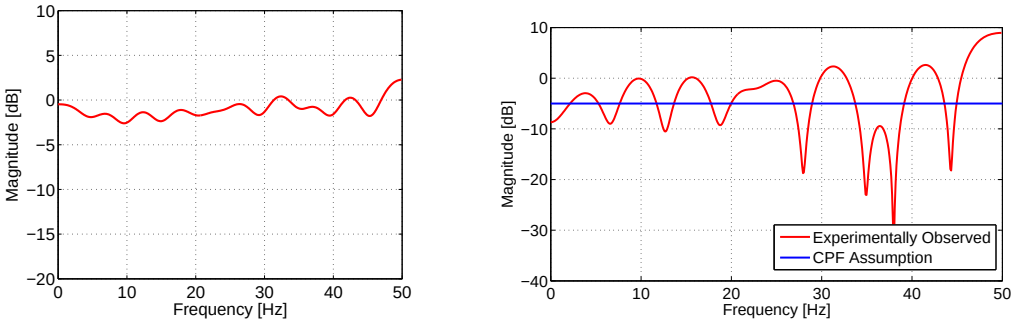
$$\phi[m] \triangleq \sum_{n=0}^{N_h-2} G_c[n]G_c[n+m] = \sum_{n=1}^{N_h-1} H_c[n]H_c[n+m] \; < \; \epsilon \; << \; 1, \; \forall m \neq 0 \qquad (4.14)$$

Thus we assume that the deterministic autocorrelation sequence of the coefficients in the filter $\mathcal{G}_c(z)$ is approximately a delta function. In Fig. 4.2(a), we plot the frequency response corresponding to an experimentally obtained filter $\mathcal{H}_c(z)$ for a real recorded RIR at an RT of 470 ms for the 0^{th} cepstral feature. We note that the frequency response of the filter $\mathcal{H}_c(z)$ is nearly flat across different frequencies. We also plot the frequency response corresponding to the autocorrelation sequence in (4.14) in Fig. 4.2(b). We note that our assumption that the frequency response is a constant, as implied by (4.14), does not strictly hold. However, the error due to the assumption is within approximately 5 dB over the modulation frequencies of 5-20 Hz, the frequency range which is believed to carry most of the speech knowledge. Thus, we experimentally support our assumption in (4.14) with respect to Fig. 4.2(b).

Note that in practice, the filter $H_c[n]$ is completely unknown and to make the dereverberation solution tractable we are required to make certain reasonable assumptions about the filter $H_c[n]$. Any additional information about the filter $H_c[n]$ in terms of its corresponding autocorrelation in (4.14) can also be incorporated in the CPF framework.

4.2.2 Error Tradeoff in Clean versus Reverberated Condition

In this section we study the consequences of CPF processing with respect to the associated error tradeoffs in the CPF framework of Fig. 4.1. Note that in the CPF framework, training is done on the W_c data sequences instead of training on the X_c sequences that correspond to the clean (unreverberated) condition. This leads to an error in clean conditions due to the differences between the X_c and W_c sequences. The CPF error in the presence of reverberation is the residual difference between the W_c and Z_c sequences. Figure 4.3 describes the tradeoff in CPF processing with respect to normalized errors in the clean and reverberated conditions. The only parameter in CPF processing is the N_p parameter that specifies the number of filter taps in the filter $P[n]$ in Fig. 4.1. We see in Fig. 4.3 that increasing the N_p parameter helps reduce the error in the presence of reverberation for all the cepstral features but simultaneously increases



(a) (b)

Fig. 4.2: (a) The frequency response for the filter $\mathcal{H}_c(z)$ for a real recorded RIR at an RT of 470 ms for the 0^{th} cepstral feature (b) The experimentally observed and the assumed frequency response corresponding to the autocorrelation sequence $\phi[m]$ in (4.14) for the filter in (a).

the error in clean condition. Thus, there exists a tradeoff between performance in the clean and reverberated conditions. For ASR applications, we experimentally find that $N_p \in \{5, 7\}$ works.

4.3 EXPERIMENTS AND RESULTS

We applied our post-filtering compensation to a subset of the YOHO database [55] for the task of speaker identification (SID) [56–59]. Reverberant speech was obtained by convolving clean speech with simulated room impulse responses produced by the *RIR* [6] simulator for room acoustics. We used a simulated room with dimensions $5 \times 4 \times 3$ m with a single microphone located at the center of the room, and a distance of 1 m between the source and the microphone. SID accuracy results using CPF are summarized in Fig. 4.4. The SID accuracy curve labeled "GMM" corresponds to a Gaussian Mixture Model based SID system, also the *uncompensated* case. "GMM-P-n" refers to the use of a post-processing filter $P[n]$ with the parameter n denoting the duration of the FIR impulse response.

Comparing the results we note that our post-filtering compensation approach provides substantial improvement in SID accuracy, with greatest improvements observed

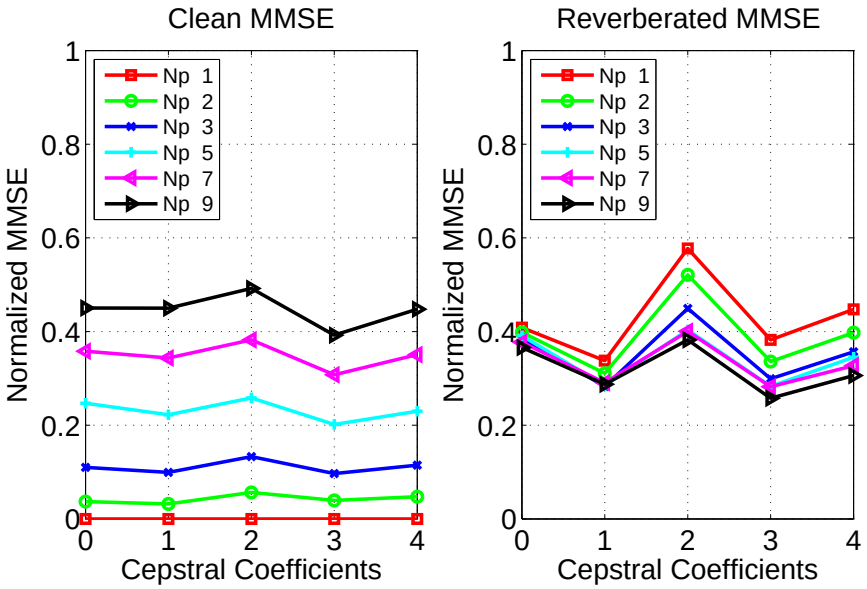


Fig. 4.3: *Error tradeoff for clean vs. reverberated condition in CPF processing.*

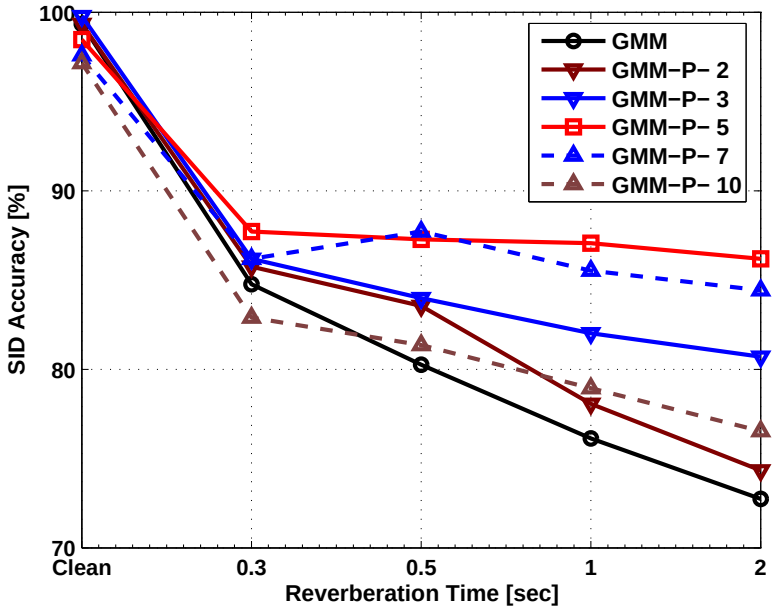


Fig. 4.4: *SID accuracy in reverberation.*

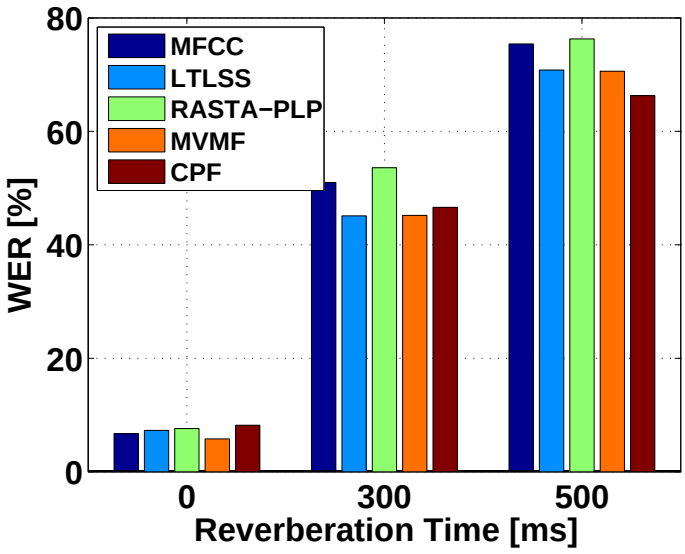


Fig. 4.5: *CPF for reverberation compensation in an ASR task.*

for the larger RTs. Best performance was obtained for the relatively small number of five filter taps, in which case the relative SID average error rate decreased by 38% compared to the uncompensated case.

4.3.1 ASR Experimental Setup and Results

The ASR system we used for training and decoding speech was the CMU Sphinx-3 [2] open source system. We trained the system on speech from clean condition and tested its performance on clean as well reverberant environments. The test data were different from the training data in all the cases. The acoustic models were all 3-state left-toright Bakis topology hidden markov models (HMMs) [60,61] with no skips permitted between states. Each state output distribution was modeled by a mixture of 8 Gaussians. The total number of tied states used was 1000. The language model used was a standard bigram model for the task under consideration, built inhouse using the CMU Language modeling toolkit. The features used were conventional 13-dimensional MFCC features augmented by delta and doubledelta cepstra. Each full feature vector was 39-dimensional. Cepstral mean normalization (CMN) was applied in all cases.

[2]Available online at `http://cmusphinx.sourceforge.net/html/cmusphinx.php`

We tested the CPF algorithm on an ASR task using the DARPA RM [12] database available from the Linguistic Data Consortium and we present the corresponding results in Fig. 4.5. We note that CPF processing provides very competitive results for more challenging reverberant environments. Specifically at RT of 500 ms condition, CPF provides the best ASR word error rate (WER) performance. In that particular reverberation condition, CPF provides a 12% reduction in relative WER and a 6% relative reduction over the best performing MVMF algorithm.

4.4 DISCUSSION

In this chapter we developed a compensation for reverberation in terms of post-filtering the cepstral sequences. The filter parameters were obtained by formulating and solving a mean-squared objective function. In our experiments, we found the optimal number of filter taps in $P[n]$ to be in 5 to 7. In Fig. 4.3 we found that although the distortion $\mathbb{E}[d^2]$ decreases with a greater number of filter taps, the similarity between the uncompensated and compensated clean features decreases as well. That led to a tradeoff for the filter-length parameter N_p with respect to error in clean and reverberated conditions. Overall, the CPF approach provided significant improvements in SID and ASR accuracy across different reverberation conditions.

CHAPTER 5

MAXIMUM-LIKELIHOOD-BASED CEPSTRAL INVERSE FILTERING

Speech dereverberation has been addressed by a number of algorithms in the past but it still remains a challenging problem. Compensation for reverberation has been difficult because the environment reverberation characteristics are generally unknown, so the compensation algorithms must work without any *a priori* information about the reverberation parameters. Some of the past work has considered blind methods for dereverberation but the improvements have been limited, possibly due to an incorrect model of reverberation or an improper domain for reverberation compensation. Our work in this chapter serves to provide a compensation algorithm that is built on an adequate reverberation model in an appropriate domain, i.e., the speech feature domain. Further, the algorithm does not require any *a priori* information about the reverberation parameters.

In this chapter we propose a maximum-likelihood-based cepstral inverse filtering (max-LIFE)[1] for compensating reverberation in the cepstral domain. Note that we have already proposed the CPF algorithm for reverberation compensation in Chapter 4. Both CPF and LIFE operate in the framework of our proposed reverberation model in Chapter 3 and design a reverberation-compensate filter to compensate for the environment filter as modeled by $H_c(z)$ in Fig. 5.1. The key difference between CPF and LIFE lies in the reverberation-compensation problem formulation. While CPF made simplified assumptions on the filter $H_c(z)$, the LIFE processing does not make any assumption about the form of $H_c(z)$. The filter is thus allowed to have arbitrary frequency response in the LIFE framework, so the LIFE approach tries to solve the reverberation compensation in a more general fashion than was attempted in the

[1]This work was published in [62].

CPF approach. The LIFE approach sets up a likelihood objective criterion to guide the reverberated features to a domain corresponding of the clean features. Likelihood based objectives have also been applied for other signal processing applications in [63]. A key advantage of the likelihood objective is that the objective does not require any *a priori* information about the reverberation filter parameters. Thus, LIFE filtering is blind to the reverberation parameters and can work across different reverberation conditions.

5.1 MOTIVATION FOR THE MAXIMUM LIKELIHOOD CRITERION FOR ESTIMATING INVERSE FILTERS

In this Chapter we seek to improve the robustness of ASR systems with respect to reverberation. In Chapter 3, we provided a model of reverberation in the cepstral domain where the model included a finite impulse response (FIR) linear time-invariant (LTI) system. Using the FIR representation of reverberation, a typical approach for reverberation compensation is to design a system which acts as an inverse for the reverberating LTI system. Nevertheless, the design of such an inverse system is difficult because the time domain reverberation filter is generally both unknown and potentially non-invertible. In this work, we propose the estimation of an inverse system to compensate for reverberation using a maximum likelihood (ML) criterion [63]. ML only requires knowledge of the probability density function (pdf) from which the signals are drawn, which can be obtained from a small amount of training data. ML transforms the reverberated signals into a space from which the clean signals are believed to originate and thus dereverberates the signal.

We first demonstrate analytically the merit of the ML criterion through two simple illustrations. We show that under certain assumptions about the original signal, it is possible to estimate approximately the optimal inverse LTI parameters from the recordings of reverberant input signals. We demonstrate our approach for both all-zero and all-pole inverse systems.

5.1.1 Inverse FIR Filter

In this illustration, we formulate a simple reverberation problem and demonstrate that we can invert the effects of reverberation with an estimated all-zero filter. We assume that the reverberated signal $x[n]$ can be represented in terms of a convolution of the original unobserved signal $s[n]$ and an FIR filter $H(z)$, which models the reverberation. We further assume that the the original signal $s[n]$ is white and Gaussian, with zero mean and an autocorrelation that is the Kronecker delta function. We assume that the filter $H(z)$ has only two taps (and hence only a single delay tap). The assumption that the reverberation filter is only of length 2 appears very restrictive at first but we can overcome this restriction by applying the approach in each of multiple narrow sub-bands. We assume that a large number of narrow sub-band versions of the H filter can approximate the actual H filter. Note that these assumptions are only for the present illustrations; we will allow the number of filter taps to be unconstrained in the actual ASR problem. We formalize our assumptions as follows:

$$\begin{aligned} s[n] &\sim N(0,1), \text{ the original signal} \\ H(z) &= 1 + h\, z^{-1}, \text{ the reverberation filter} \\ x[n] &= s[n] * h[n] = s[n] + h\, s[n-1], \text{ the reverberated signal} \end{aligned} \tag{5.1}$$

Next we formulate our problem in terms of designing a filter that operates on the reverberated signal $x[n]$ using a log-likelihood criterion with respect to the pdf of $s[n]$, to design the inverse filter parameters. In Eq. (5.2) below, $P(z)$ denotes the supposed inverse FIR filter and $y[n]$ is the estimated dereverberated signal.

$$\begin{aligned} P(z) &= 1 + p\, z^{-1} \\ y[n] &= x[n] * p[n] = x[n] + p\, x[n-1] \end{aligned} \tag{5.2}$$

The filter parameter p is estimated by maximizing L, the likelihood of $y[n]$ with respect to the pdf of $s[n]$.

$$L = \log \Pi_n \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2[n]}{2}\right) \tag{5.3}$$

The above can be simplified to minimizing

$$L = E[y^2[n]] \tag{5.4}$$

where we replaced summation by expectation and ignored positive constants under the operation. The optimal filter parameter is obtained by differentiating L with respect

to the unknown p.

$$\begin{aligned}\frac{\partial L}{\partial p} &= 2E[(x[n] + p\,x[n-1])x[n-1]] \\ &= 2(R_{xx}[1] + p\,R_{xx}[0])\end{aligned}$$

Setting the above to 0, we obtain:

$$p = -R_{xx}[1]/R_{xx}[0]$$

Noting that $x[n]$ is the convolution of $s[n]$ and filter $H(z)$, the relationship between the autocorrelation sequence of $x[n]$ and $s[n]$ becomes

$$R_{xx}[n] = R_{ss}[n] * R_{hh}[n] \tag{5.5}$$

It can easily be shown that

$$R_{xx}[n] = [h,\ 1+h^2,\ h], \quad n \in [-1, 0, 1] \tag{5.6}$$

from which we obtain

$$p = -h/(1 + h^2)$$

Next, assuming that $h \ll 1$ and making the first-order approximation of neglecting the squared term for h, we obtain

$$p \approx -h \tag{5.7}$$

The assumption of $h \ll 1$ holds if we work on narrow sub-bands of H and assume that the frequency response of H may be nearly a constant for each of the sub-bands. So far, we showed that under the assumptions in Eq. (6.1), we can devise a log-likelihood criterion to invert the filter H by $P = [1 \ \ -h]$, which indeed is expected to be the inverse of filter H under first-order approximations. Finally we note that Eq.(6.1) does not include an explicit gain term for the reverberation filter H. While the maximum likelihood procedure described cannot be used to estimate the gain term, gain inversion can be achieved via variance normalization.

5.1.2 Inverse IIR Filter

In the illustration in Sec. 5.1.1 we showed that we can estimate an inverse reverberation filter in terms of a FIR filter. In this illustration, we start with the same assumptions

as in Eq. (6.1) but we model the inverse filter as an all-pole IIR filter, showing that we can approximately estimate the optimal inverse filter parameters. Specifically, we assume that the inverse filter P and the dereverberated signal $y[n]$ are of the form:

$$\begin{aligned} P(z) &= \frac{1}{1 + p\, z^{-1}} \\ y[n] &= x[n] * p[n] = x[n] - p\, y[n-1] \end{aligned} \tag{5.8}$$

Following the same principles as in Sec. 5.1.1 we obtain

$$\frac{\partial L}{\partial p} = 2E[(x[n] - p\, y[n-1])y[n-1]]$$

As before, it can be shown that $p = R_{xy}[1]/R_{yy}[0]$ and

$$\begin{aligned} R_{xy}[1] &= R_{xx}[1] = h \\ R_{yy}[0] &= \frac{1 + h^2 - 2p\, h}{1 - p^2} \end{aligned}$$

from which we obtain:

$$p = \frac{h(1 - p^2)}{1 + h^2 - 2p\, h}$$

Clearly, $p = h$ is one of the two solutions in the above quadratic equation and the estimated compensation filter becomes:

$$P(z) = 1/(1 + hz^{-1})$$

which is indeed the inverse of the filter $H(z)$.

Sections 5.1.1 and 5.1.2 illustrated the maximum likelihood formulation for estimating a filter (FIR or IIR) that inverts the effects of reverberation. The illustrations showed analytically that our approach is well founded and can approximately guarantee the optimal performance under certain assumptions. While these assumptions, of course, do not hold for speech signals, we relax some of those assumptions in Sec. 5.2 which follows, and we extend the approach to ASR. While analytical verification of our approach for realistic reverberant environments is not tractable, we validate our approach through experimental results in Sec. 5.4.

5.2 MAXIMUM-LIKELIHOOD-BASED INVERSE FILTERING (MAX LIFE)

In Sec. 5.1 we formulated the problem of reverberation compensation in terms of obtaining an appropriate inverse filter, proposing the use of a maximum likelihood criterion for obtaining that inverse filter. We demonstrated that the approach can approximately estimate the optimal inverse filter parameters. In the present section we extend our approach for reverberation compensation for speech data, referring to the extended approach as *maximum likelihood based inverse filtering* (Max-LIFE). We build our current work on the cepstral domain reverberation model derived in Chapter 3. This aforesaid model characterized reverberation as linear filtering in the cepstral domain. Some other recent dereverberation work is in a similar framework including *e.g.* [19, 54, 64, 65]. Our reverberation model extends the earlier representations of reverberation as a simple additive shift in the log-spectral or cepstral domains. Continuing along these lines we seek to design an inverse reverberation filter that does not make any *a priori* assumptions about the nature of the actual room reverberation filter. We formulate a maximum likelihood objective function which requires the pdfs of the features of clean speech, which can be obtained from training data. The likelihood objective is expected to guide the features to the space from which the clean features originate, thereby dereverberating the features. In Sec. 5.1.1 the pdf was assumed to be a single Gaussian density in Eq. (6.1). Since a single Gaussian density is insufficient for real speech applications, we extend the pdf to be a Gaussian mixture model (GMM), trained from a pool of clean speech features. The number of filter taps to model reverberation was assumed to be 2, in the discussion of Sec. 5.1.1, but for practical ASR the number of filter taps modeling reverberation will need to be unconstrained.

5.2.1 Mathematical Formulation of LIFE Filters

We show the working of LIFE algorithm in Fig. 5.1, which summarizes the processing for a unidimensional feature. The approach can easily be extended to multidimensional features by individually applying the processing to each of the multidimensional features under the assumption that the features are uncorrelated. The $X_c[n]$ in Fig. 5.1 is a unidimensional cepstral feature for clean speech, whereas $Y_c[n]$

models the corresponding reverberated feature. The model was derived in Chapter 3. The LIFE processing applies a two-fold compensation scheme to $Y_c[n]$. It first compensates for the additive constant C_Δ by mean normalization as is also done in CMN. The effects of $H_c(z)$ are normalized by a $P(z)$ filter.

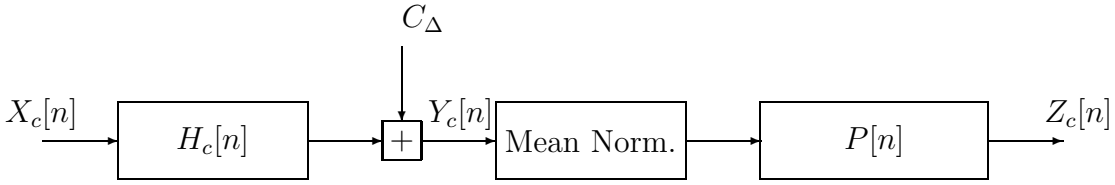


Fig. 5.1: *LIFE compensation in cepstral feature domain.*

The $P(z)$ filter parameters are designed by a maximum-likelihood (ML) criterion. In general the ML criterion is based on a probability distribution which in our work corresponds to the distribution of clean features in $X_c[n]$. We assume that the features $X_c[n]$ are distributed according to a Gaussian Mixture Model (GMM) and we learn the GMM parameters from available clean training data. Among other choices, the $P(z)$ filter can be chosen to either be a finite impulse response (FIR) or an all-pole infinite impulse response (IIR) filter. While we illustrate these developments only for the inverse IIR filter, the approach can easily be adapted for the FIR filters. We derive the updated equations assuming $P(z)$ to be an all-pole IIR filter with M coefficients, the reverberation-compensated features ($Z_c[n]$) become:

$$Z_c[n] = Y_c[n] - \sum_{m=1}^{M-1} p[m] Z_c[n-m] \tag{5.9}$$

The parameters that describe P are obtained by maximizing the log-likelihood with respect to the GMMs for speech. Specifically, $\mathbf{P} = \arg\max_P L$, where the log-likelihood L for the compensated features is

$$L = \frac{1}{N_z} \sum_{j=1}^{N_z} log \left(\sum_i \frac{w_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(Z_c[j] - \mu_i)^2}{2\sigma_i^2} \right) \right) \tag{5.10}$$

The GMM parameters are represented by the set $\{w_i, \mu_i, \sigma_i\}$ with N_w being the number of Gaussian densities and N_z being number of feature frames in $Z_c[n]$. For the ease of

writing and understanding the equations we define:

$$\begin{aligned}\gamma_i^j &= \frac{w_i}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(Z_c[j]-\mu_i)^2}{2\sigma_i^2}\right) \\ \gamma^j &= \sum_{i=1}^{N_w} \gamma_i^j \\ L &= \frac{1}{N_z}\sum_{j=1}^{N_z} \gamma^j\end{aligned} \tag{5.11}$$

Using the above definitions we maximize Eq. (5.10) by gradient ascent via its partial derivative with respect to the parameters in P. It can be shown that:

$$\frac{\partial L}{\partial p[m]} = \frac{1}{N_z}\sum_{j=1}^{N_z}\sum_{i=1}^{N_w} \frac{\gamma_i^j}{\gamma^j} \; \frac{(Z_c[j]-\mu_i)\; Z_c[j-m]}{\sigma_i^2} \tag{5.12}$$

Next, we iteratively obtain the parameters for $p[m]$:

$$\hat{p}[m] = p[m] + \nu \frac{\partial L}{\partial p[m]} \tag{5.13}$$

where ν is a small-valued learning-rate parameter. The filter update in (5.12) provide a deep understanding into the evolution of P. Summing over the j terms for a fixed i in Eq. (5.12) results in an update for $p[m]$ that is proportional to the m^{th} auto-correlation sequence of $Z_c[n]$. Summing over the i terms for a fixed j in Eq. (5.12) results in $p[m]$ becoming proportional to the summed and weighted likelihoods of γ_j^i. Thus the overall filter updates are proportional to the "likelihood-weighted" auto-correlation sequences of $Z_c[n]$.

Note that Eq. (5.12) requires knowledge of $Z_c[j]$ which in turn depends on P in Eq. (5.9), so $Z_c[j]$ will also be updated after each iteration of P.

5.2.2 The Top-1 Approximation for Filter Updates

The filter update described in Eq. (5.13) may be simplified through suitable approximations. A common approximation in GMMs is to replace the overall GMM likelihood score in Eq. (5.10) by the top-scoring Gaussian density among the set of Gaussian

mixtures. This approximation, referred to as the Top-1 approximation, results in:

$$\begin{aligned} \gamma^j &= \sum_{i=1}^{N_w} \gamma_i^j \approx \gamma_{i_*}^j, \quad i_* = \arg\max_i \gamma_i^j \\ \frac{\partial L}{\partial p[m]} &= \frac{1}{N_z} \sum_{j=1}^{N_z} \frac{(Z_c[j] - \mu_{i_*})\, Z_c[j-m]}{\sigma_{i_*}^2} \end{aligned} \tag{5.14}$$

Note that i_* is a function of j in Eq. (5.14). This approximation is more valid for sparsely-distributed features in terms of the Gaussian densities where only the top-scoring density can adequately describe the overall feature score. A Top-N approximation could be similarly derived by approximating Eq. (5.10) with the top-N Gaussians.

5.3 LIFE FILTER PARAMETERS

In this section we finalize a LIFE processing framework with a robust set of parameters determined experimentally. Please see Sec. 4.3.1 for details on the ASR experimental setup.

5.3.1 Pre-processing for LIFE Filters

The LIFE processing framework assumes the cepstral features to be independent across time but in practice the features are correlated and hence not independent. Considering this, a pre-processing step before the LIFE filtering stage that may partially decorrelate the features may assist the LIFE processing. In Chapter 4 we noted that CPF processing has an effect of partially decorrelating the features, consequently we experimented with pre-processing the LIFE filters with CPF processing. We present the corresponding ASR WER results on DARPA-RM [12] database in Fig. 5.2. We see that at RT of 300 ms, CPF provides a relative reduction of 8.6% in WER compared to MFCC. LIFE processing without CPF pre-processing provides a 31.3% relative reduction in WER over MFCC but LIFE with CPF pre-processing the features provides 40% relative reduction, thus the CPF pre-processing strongly assists LIFE processing. Interestingly we note that CPF processing improves the WER obtained using MFCC by 8.6% while it improves the processing with LIFE by 13.7%. In other words CPF works

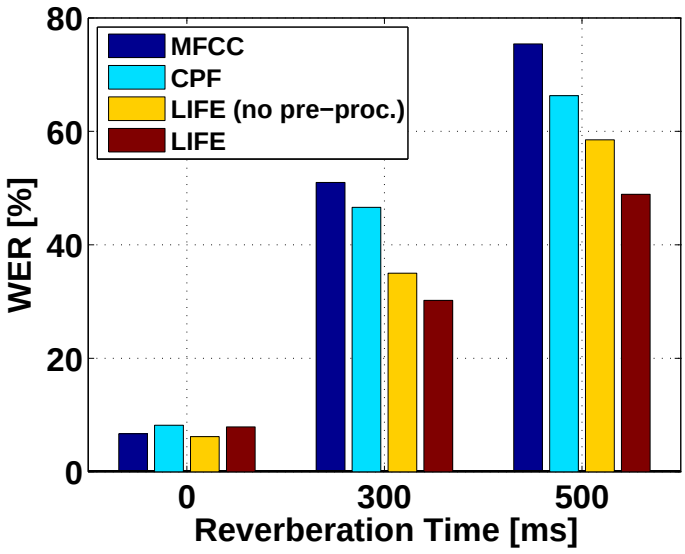


Fig. 5.2: *A pre-processing stage for LIFE filter.*

in strong synergy with LIFE processing, so we include a CPF pre-processing stage, where the cepstral features are first passed through a CPF stage and the resulting cepstral features are then passed through LIFE processing.

In the above experiments, LIFE processing was applied with a single-density Gaussian model to model the speech cepstral features. Next, we experimentally evaluate LIFE processing for multiple Gaussian densities.

5.3.2 Number of Gaussian densities in Modeling Speech Cepstral Features in LIFE Processing

In Fig. 5.3 we present our LIFE-based reverberation compensation results for multiple Gaussian densities in modeling the speech cepstral features. We see that increasing the number of Gaussian densities does not substantially improve the WER. We also highlight an interesting tradeoff associated with increasing the number of Gaussian densities. The advantage includes a better modeling for the overall distribution of the speech feature. But this also results in a poorer model for individual utterances and since LIFE processing works on an utterance basis, where the utterances are typically 4-5 seconds long, increasing the Gaussian densities offers a modeling tradeoff. Increasing

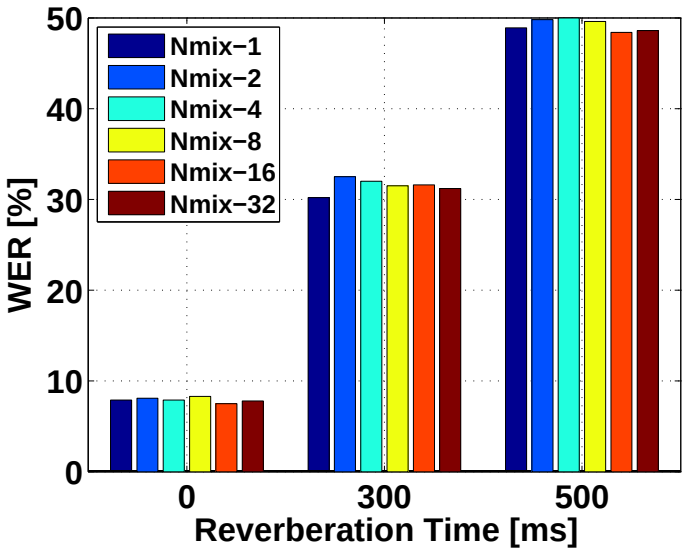


Fig. 5.3: *LIFE processing for different Gaussian densities.*

Gaussian densities also increases the computational complexity of the LIFE algorithm. On the basis of the results in Fig. 5.3 we determine the number of Gaussian densities to be 1 in subsequent implementations of LIFE framework.

5.3.3 Duration of Filter Taps in LIFE Processing

The number of filter taps *i.e.* M in (5.9) is an important parameter in the LIFE algorithm. In Fig. 5.4 we evaluate the LIFE algorithm for different number of filter taps. Note that the frame-shift interval in the feature extraction was 10 ms, so each filter-tap equivalently spans a duration of 10 ms. We see that 20 taps, corresponding to 200 ms, works nearly the best for LIFE processing at different reverberant conditions.

5.3.4 An Oracle Experiment with Zero Modeling Error

LIFE processing works on the reverberation model derived in Sec. 3.3.1. There, model derivation incurred an additive error term that we ignored in our final model in Fig. 3.10. In practice, the model will have some errors and may degrade the overall dereverberation performance of LIFE processing. We conducted an oracle experiment

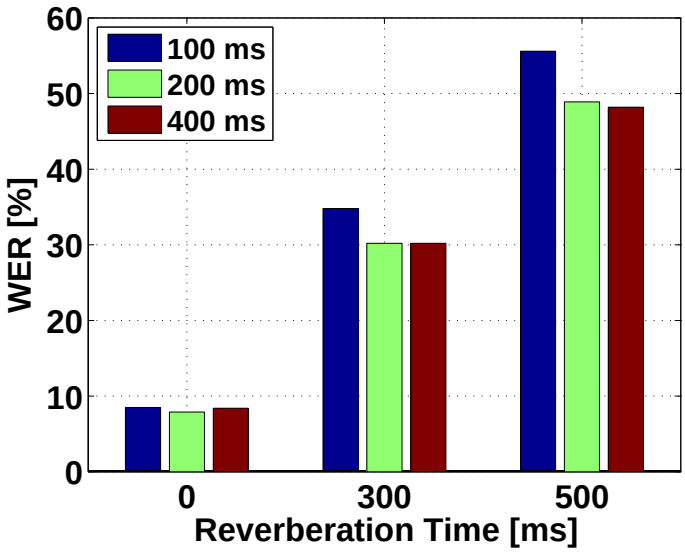


Fig. 5.4: *LIFE processing for different filter lengths. A single filter-tap spans 10 ms for a feature extraction scheme with the frame sampling frequency of 100 Hz.*

to test LIFE processing under zero modeling error. This oracle experiment isolates and identifies the potential merits of LIFE processing that would be obtained with a correct model. We simulated the oracle experiment by directly convolving the cepstral features with exponentially decaying RIRs corresponding to RTs of 300 and 500 ms. LIFE processing was applied to the convolved features. Note that even though the modeling error was zero, LIFE processing is still blind to the actual RIR. We present the corresponding ASR results in Fig. 5.5. LIFE processing with zero modeling error achieves a WER of 11% at the simulated RT of 500 ms, which is fairly close to the MFCC baseline for clean conditions of 6.7%. Thus we conclude that LIFE processing by itself has strong merit and can provide huge improvements in absence of reverberation modeling error. Finally, comparing WER in the above oracle experiment with that obtained in practical reverberation environments in Fig. 5.2, we note that although LIFE processing provides substantial improvement in WER, its full potential is limited due to modeling error. We hope that a future improvement that reduces modeling error before LIFE processing can provide substantial further improvement for ASR.

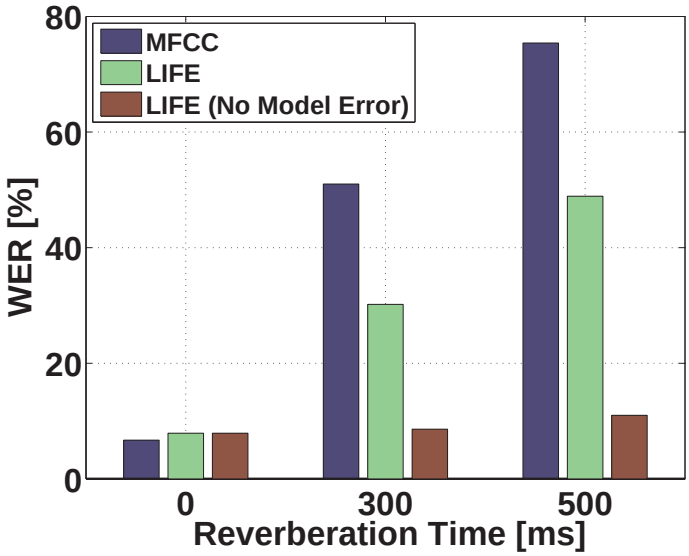


Fig. 5.5: *LIFE processing for an oracle experiment with zero modeling error.*

5.4 DATABASES AND RESULTS

In this section we experimentally evaluate the LIFE algorithm against some of the established baseline and reverberation compensation algorithms. Fig. 5.6 summarizes our ASR dereverberation experimental results using the DARPA RM database. There, the LIFE GMMs were trained using a single Gaussian density and the inverse P filter was 20 taps long. We note that compared to the MFCC features, the baseline algorithms in LTLSS, RASTA-PLP, MVMF, CPF provide only about a 10-15% relative reduction in WER across different reverberation conditions, but LIFE processing provides 35-40% relative reduction in WER compared to results using MFCC features.

5.4.1 Composite LIFE Filter

In Fig. 5.7 we plot the frequency responses corresponding to the estimated LIFE filters for the "C1" cepstral feature for 4 different utterances from the RM database, the utterances all having been recorded at RT of 300 ms. LIFE filters were individually obtained for each of the utterances. We see that the frequency responses exhibit a high-pass characteristics, which is expected because reverberation smears the speech spectrum and acts like a low-pass filter in time. Consequently the dereverberation

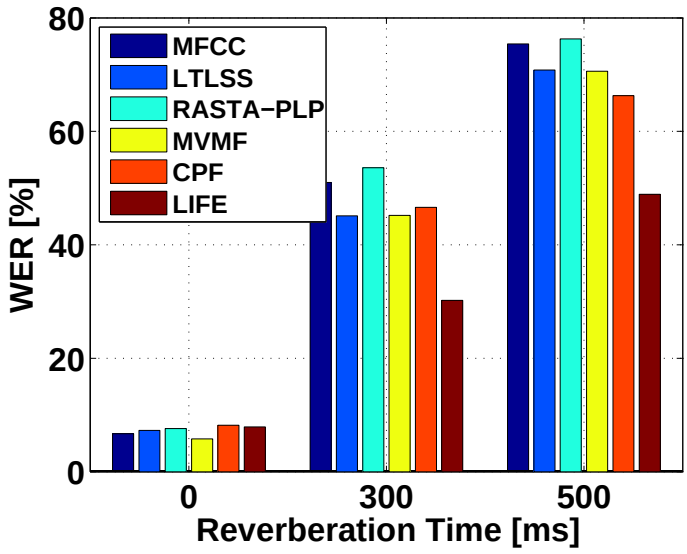


Fig. 5.6: *WER comparisons for LIFE processing.*

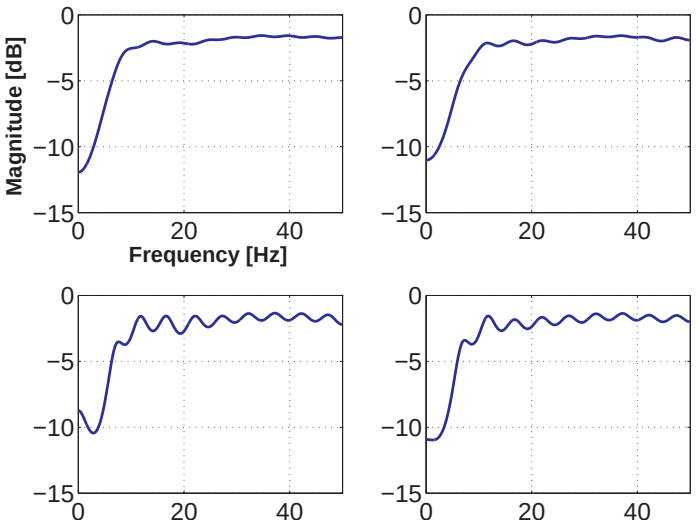


Fig. 5.7: *LIFE frequency responses for the "C1" cepstral feature for 4 different utterances at RT of 300 ms.*

filter is expected to exhibit a high-pass characteristic.

The LIFE filters for the different utterances for a particular reverberation condition are similar but not identical. We ideally expect the LIFE filters to be identical for

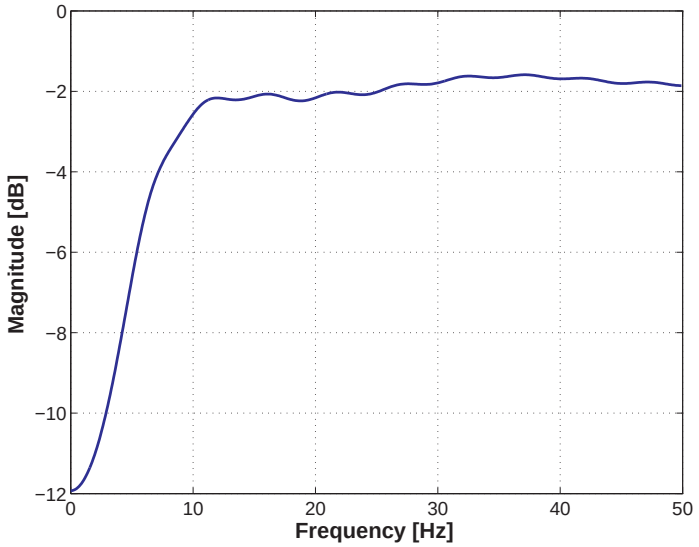


Fig. 5.8: *The average frequency response for the "C1" cepstral feature at RT of 300 ms.*

a particular reverberation condition. This raises an important question as to whether the differences among the filters are significant for ASR. Alternatively, is it necessary for the LIFE filters to be derived on an utterance-by-utterance basis? If not, could a composite-LIFE filter be obtained by averaging over the LIFE filters for different utterances for a particular reverberation condition? To answer the above questions, we conducted an experiment in which for each reverberation condition, we individually obtained the LIFE filters for 20 different utterances belonging to that particular condition. We also obtained a composite-LIFE filter by averaging over the 20 individual LIFE filters and appropriately applied this composite-LIFE filter to all the utterances belonging to that particular reverberation condition in the database. Note that the LIFE filters are different for the different reverberation conditions but are identical for the different utterances of a particular reverberation condition. We expect the composite-LIFE filters to work similar to the original LIFE processing.

In Fig. 5.8, we plot the frequency response of the composite LIFE filter at RT of 300 ms and in Fig. 5.9, we compare the LIFE and composite-LIFE algorithms. These results verify that composite-LIFE filters perform similar to the original LIFE filter, and hence we note that the an individual LIFE filter for a particular reverberation

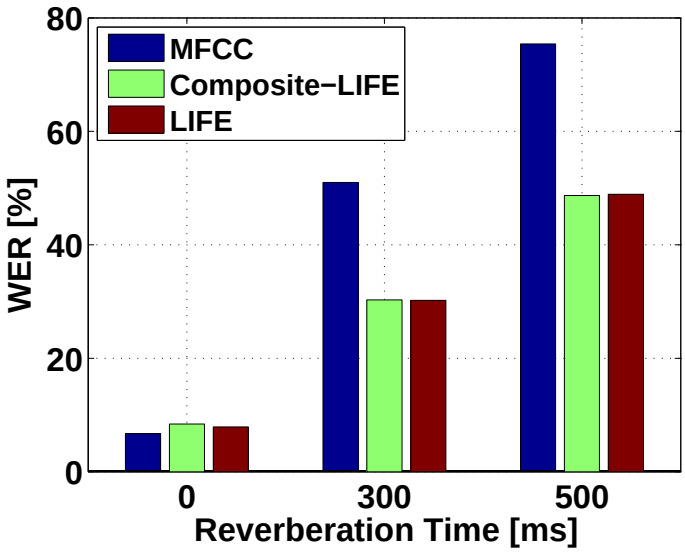


Fig. 5.9: *WER for the composite-LIFE filter that was evaluated for each of the different room conditions.*

condition incorporates a strong average characteristics which is strongly dependent upon the reverberation condition.

5.4.2 Experiments on Real Room Impulse Response

In all the experiments so far, the reverberant speech was obtained by convolving clean speech with a simulated room impulse response (RIR). In this section, we apply LIFE processing on data obtained from a real RIR. We obtained an RIR corresponding to RT of 500 ms from the ATR [13] database. In Fig. 5.10, we plot the WER for MFCC and LIFE processing. We see that benefits from LIFE processing extend to real RIRs, and at RT 500 ms, we obtain 38% relative reduction in WER. This improvement favorably compares with the 35-40% relative reduction in WER for RTs 300-500 ms conditions in Fig. 5.6.

5.4.3 Experiments on Recorded Speech in Non-Stationary Environment

All of the experiments reported so far were done on speech utterances convolved with either simulated or real RIRs. This implicitly assumed the room to be stationary

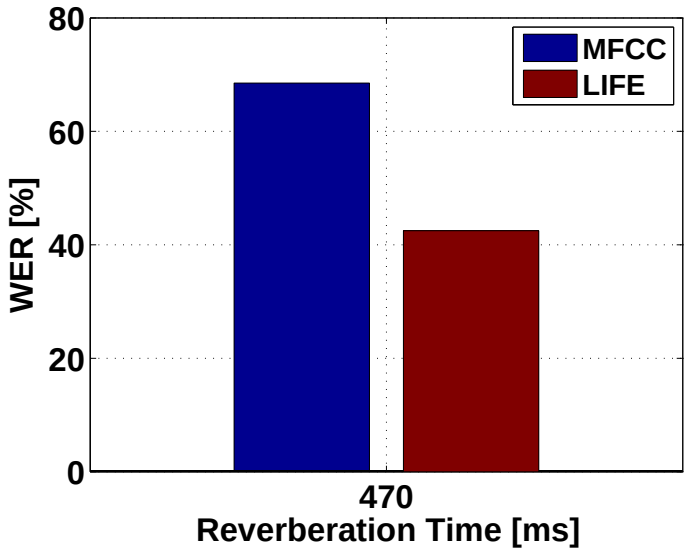


Fig. 5.10: *LIFE processing on a recorded RIR from ATR database.*

which may not be true in practical environments. Usually speaker and listeners may move around in room which leads to non-stationarities in the RIR. Although LIFE processing worked substantially well in stationary environments, it does not guarantee its performance in realistic non-stationary environments. In Fig. 5.11 we report LIFE processing results on an ATR [13] database with TIMIT utterances recorded in a room with RT of 470 ms. There the speaker was moving in circular motion and the experiment setting incorporates a non-stationary reverberant environment. We note that LIFE processing works in non-stationary environments as well and provides a 45% relative reduction in WER compared to baseline MFCC processing.

5.4.4 Multi-style Experiments

The experimental paradigm in all the results reported so far was with clean training, *i.e.*, training was done on clean speech, and decoding was done on the acoustic models obtained from the clean training. In this section, we introduce a multi-style experimental paradigm for ASR. In the multi-style paradigm, the training set includes data from multiple available environment conditions [66]. For reverberation compensation experiments, the training set may include clean data as well as data from a few reverberant conditions.

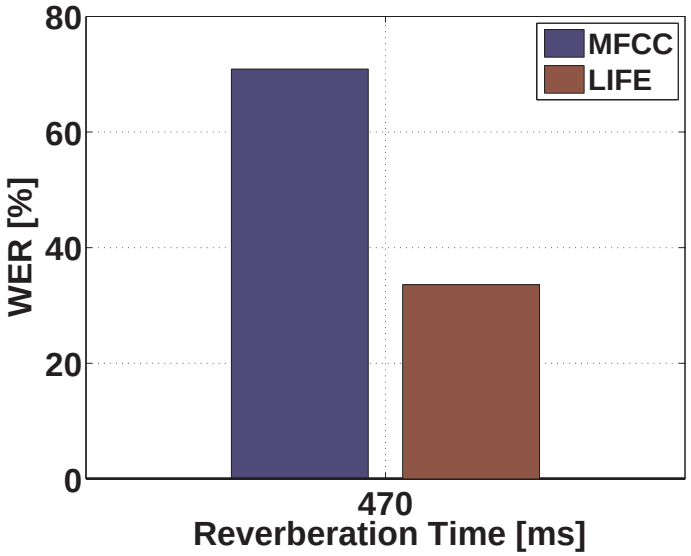


Fig. 5.11: *LIFE processing in non-stationary environments.*

The multi-style experimental paradigm has become exceedingly important for speech recognition applications, as it has been reported that a lot of robustness to noise and reverberation can be easily obtained by a simple multi-style training paradigm. Interestingly, training does not necessarily require the test condition data to be present during training. Thus, multi-style training does not require a strict match between the training and testing conditions and significantly enhances the merit in the approach.

Next we elaborate our framework for multi-style training experiments. We obtained data from 3 different environment conditions as mentioned below:

1. **R1:** Data from this environment were simulated through the RIR [6] software. The room dimensions were $5 \times 4 \times 3$ m.
2. **R2:** Data from this environment were simulated through the RIR software in above. The room dimensions were $8 \times 6 \times 4$ m.
3. **JR2:** The ATR database includes real recorded RIRs from a few different rooms. The particular data used were for a room with RT of 500 ms.

In Fig. 5.12, we present our speech recognition experiments in a multi-style setting. There, the training set data included data from the R1 environment with 3 different reverberation times of 0, 200, and 500 ms. Testing was performed across the test

conditions used in Fig. 5.12. We make the following observations about the multi-style training paradigm:

- The multi-style training paradigm by itself provides a 60-70% relative reduction in WER. This is a huge improvement over the clean training paradigm. We also note that multi-style paradigm also works when data from a particular test condition are not present in the training data. For example, the training conditions in Fig. 5.12 include data only from the environment R1, but we see that this training also provides a 64% relative reduction in WER for data from environment R2 at RT of 300 ms condition. This is an extremely promising result for the multi-style setting as it suggests that the results do not critically depend upon a strong match between the training and testing conditions and that multi-style training can work even if data from a particular room condition and/or from a particular environment is not present in the training set.
- Under clean-training paradigm, LIFE processing provides a 30-45% relative reduction in WER over MFCC features. But under the multi-style training paradigm even the MFCC features provides 60-70% relative reduction in WER, thus the clean-training-based LIFE processing is no match for the multi-style-training-based MFCC processing. Comparing MFCC and LIFE under multi-style training, we note that on average LIFE processing provides an 18% improvement in WER compared to MFCC processing. While it is slowly being found that many of conventionally popular algorithms for robust speech recognition provide little or no benefit in the multi-style training, LIFE processing provides significant benefits under multi-style training.
- As expected, the improvement from LIFE processing is different for the various test environments. Data from the R1 condition were present in the training condition, and thus provides a partial match for data from test conditions from R1 condition. In this case LIFE processing provides on average a 12% relative reduction in WER. Data from the environments R2 and JR2 were not present in the training data. For the JR2 test condition LIFE provides a 22% relative reduction in WER. Thus, the improvements from LIFE are greater for a looser match between training and test conditions.

In Fig. 5.13 we present our WER results for multi-style training framework us-

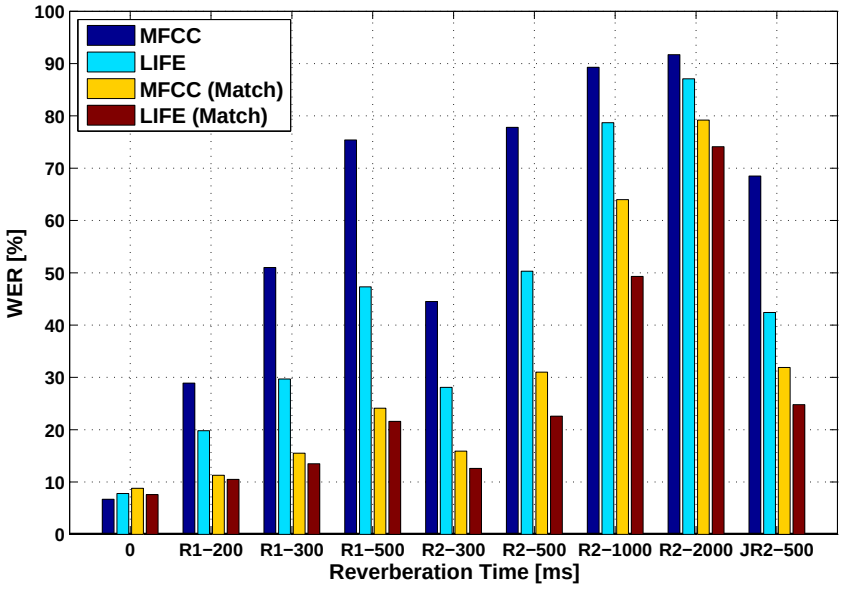


Fig. 5.12: *LIFE processing in a multi-style training paradigm.*

ing the DARPA WSJ database. There the training set included data from the R1 environment with RT of 0, 300, and 500 ms conditions. Our observations for the reverberation-compensation experiments on the WSJ database are similar to those for the RM database in Fig. 5.12. We see that for MFCC features, when testing in RT of 300 ms multi-style training provides a 62% relative reduction in WER at RT of 300 ms test condition. But we also note that under multi-style training LIFE processing provides an additional 17% relative reduction in WER over MFCC. Thus the benefits from LIFE processing extend to the multi-style training conditions on WSJ database.

In Fig. 5.14, we compare the multi-style training condition in Fig. 5.13 with a specific type of multi-style training where the training set consists of data only from the environment R1 with RT of 300 ms. Thus data from the R1 environment at RT of 300 ms will be strictly matched with data in training condition and that is expected to perform well in ASR experiments. In Fig. 5.14, we first note that the clean data (RT of 0 ms) performs very poorly for the "Match R1-300" condition. This is so because the training data did not include clean speech. For the MFCC features with RT of 300 ms, the "Match R1-300" provides a 6.5% relative reduction in WER over "Match" (matched) condition training. There, an improvement was expected for the

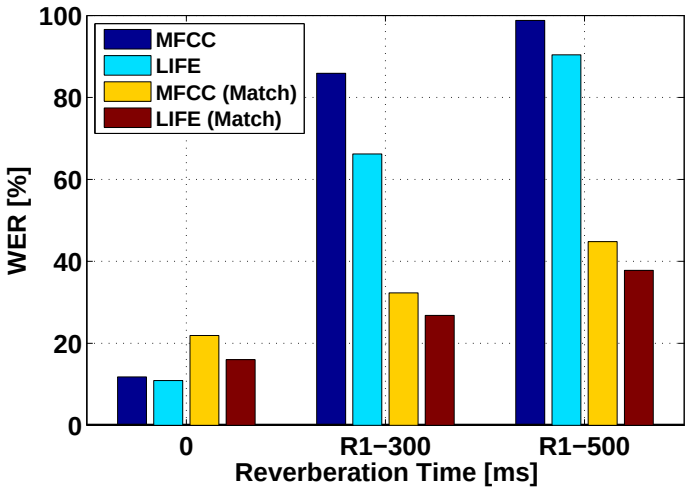


Fig. 5.13: *WER comparisons for LIFE processing on WSJ database on clean and multi-style training paradigms.*

"Match R1-300" case since data from that particular environment and only from that environment was present in the training set. Finally we also note that LIFE processing provides an improvement even when the training and test sets are strictly matched. For the "Match R1-300" test case at RT of 300 ms, LIFE provides a 18% relative reduction in WER.

5.4.5 MLLR Experiments

In this section, we study MFCC and LIFE features in the context of maximum-likelihood linear regression (MLLR) [67, 68] processing framework. MLLR applies a transformation to the means of the acoustic models to adapt the models to a particular environment condition. MLLR experiments require transcripts of the spoken utterances and correspondingly we undertake unsupervised (UnSup) and supervised (Sup) versions of MLLR. In the unsupervised setting, the transcripts are obtained by decoding on the original acoustic models, and these transcripts are then used for MLLR computation. These transcripts will in general be erroneous and may not result in good MLLR performance. In supervised experiments, the transcripts are assumed to be known *a priori*, which should provide better performance than unsupervised MLLR.

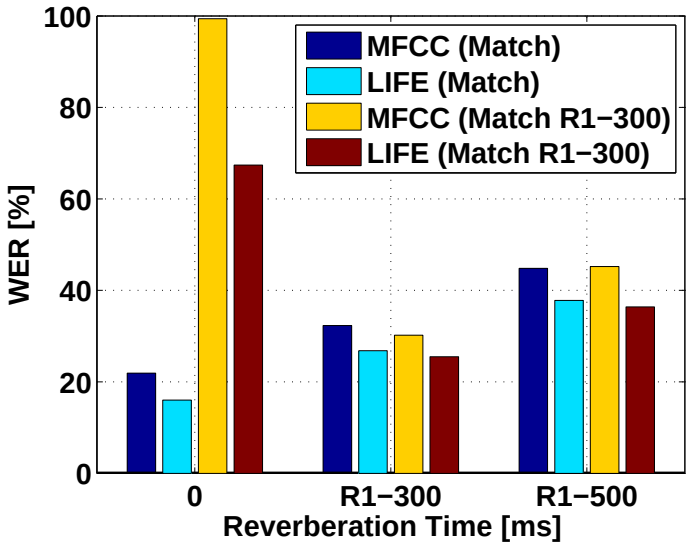


Fig. 5.14: *WER comparisons for LIFE filter under a strongly matched condition at RT of 300 ms.*

In Fig. 5.15, we compare MFCC and LIFE in the context of MLLR on RM database. We see that adding unsupervised MLLR to MFCC provides a 17% relative reduction in WER at RT of 300 ms condition, while the corresponding supervised MLLR provides a 50% relative reduction in WER. Comparing LIFE with unsupervised MLLR, we note that LIFE (wihout any MLLR) provides 40% relative reduction in WER, which is far better than the unsupervised MLLR using MFCC features. Finally, comparing supervised MLLR for MFCC and LIFE features, we see that LIFE processing provides an additional 14% relative reduction in WER. Thus the benefits of LIFE processing extend processing that includes MLLR.

5.5 DISCUSSION

In this chapter we considered the problem of dereverberation for ASR. We proposed an algorithm based on a new framework for studying the problem of reverberation for ASR. Based on our model, we motivated and developed a maximum-likelihood-based inverse filtering technique for dereverberation. The LIFE filter parameter estimation does not require any *a priori* information about the room conditions and thus is

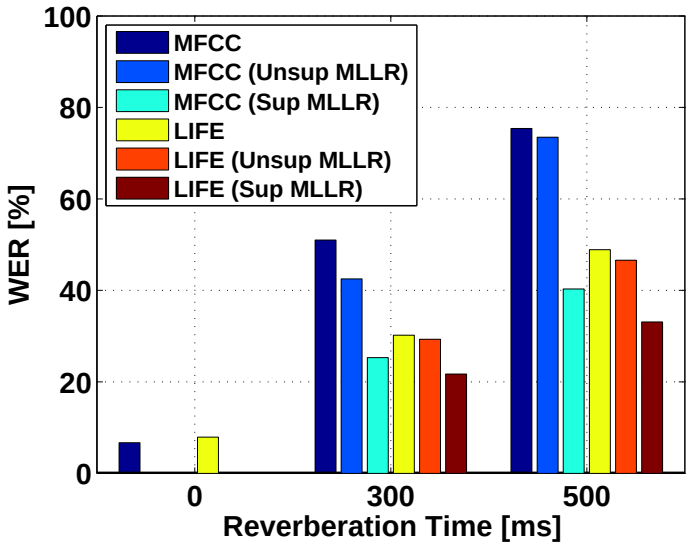


Fig. 5.15: *WER comparisons for LIFE filter.*

blind to the actual room response. The filters parameters are derived from a short speech utterance (4-5 s). Thus, the filters will in general be different across utterances and also different across reverberation conditions. Since LIFE filters compensate for reverberation, it is expected that the differences in filter parameters will primarily be due to the different reverberation conditions and not due to differences in the speech utterances. We conducted an experiment to verify this hypothesis and found that the composite filter worked nearly as well as the original LIFE filter. We demonstrated our approach in reverberant environments, and compared to baseline features we obtained up to 35-40% relative reduction in WER.

CHAPTER 6

NON-NEGATIVE MATRIX FACTORIZATION FOR SPEECH DEREVERBERATION

In this chapter we propose a spectral-domain based algorithm for dereverberation. Our work builds on the model of reverberation in the spectral domain derived in Chapter 3. For convenience we show the model in Fig. 6.1. There, $X_s[n]$ represents the spectral components corresponding to clean speech for the discrete time index n. $H_s[n]$ filter encodes the effects of reverberation and $Y_s[n]$ is the observed spectral value. The model thus represents reverberation in the spectral domain as a convolution operation between the underlying clean spectral components $X_s[n]$ and the filter $H_s[n]$. The algorithm utilizes the property that spectra are non-negative, and uses this non-negativity as a constraint to factor the observed reverberated spectra into individual components corresponding to X_s and H_s. This algorithm is called non-negative matrix factorization (NMF) [1].

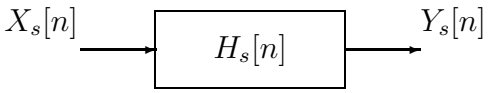


Fig. 6.1: *Modeling reverberation in the spectral feature domain.*

6.1 MATHEMATICAL FORMULATION OF NMF

Our approach for dereverberation using the spectral domain model presented in Sec. 3.2 is to try to estimate the spectrum of clean speech X_s through a decomposition of the reverberated speech spectrum Y_s into its convolutional components X_s and H_s. In this section we formulate a least-squares error criterion to achieve this decomposition.

[1] This work was published in [69].

In general, reverberation compensation algorithms should not require *a priori* knowledge of nature of the reverberation. This is the case for our algorithm as well. We do not require any knowledge of X_s and H_s. Our model of reverberation represents the effects of reverberation as the filter H_s, which is not observed directly. Rather we attempt to infer the parameters of H_s through the reverberated spectrum Y_s. This problem is highly unconstrained, and there exist infinitely many decompositions of Y_s into X_s and H_s. To constrain the solution space, it becomes necessary to assume some knowledge about either X_s or H_s that we can use as constraints. In our work, we choose two such constraints. One is that the spectral components are non-negative *i.e.* all the elements in X_s and H_s are greater than or equal to 0. This is apparent since the magnitude spectra are inherently non-negative. The second assumption is an optional one, wherein we assume that the clean spectra X_s are sparse. Later in this chapter we discuss these constraints in greater detail.

To solve the problem of decomposition we use a non-negative matrix factorization (NMF) framework. NMF was initially proposed for data clustering application in [70]. It was further developed and applied to audio applications in [71, 72], and for speech signal dereverberation in [64]. We use the NMF paradigm in [64] [71] to build our framework for dereverberation for ASR. The NMF-based dereverberation work in [64] was guided for speech enhancement applications, whereas, our work in this chapter focuses on improving ASR performance. A key contribution of our work over [64] is the incorporation of Gammatone filters [73] in the NMF framework.

Next we consider the mathematical formulation of NMF. We first assume that our actual observation sequence is $Z_s[n,k]$, which is approximately $Y_s[n,k]$

$$Z_s[n,k] \approx Y_s[n,k] = X_s[n,k] * H_s[n,k] \tag{6.1}$$

The differences between Z_s and Y_s can result from observation noise or from errors in decomposing Z_s into the convolutional components X_s and H_s. Using (6.1), we define our objective to be the minimization of the mean-squared error between Z_s and Y_s. This objective function is minimized by a gradient descent process that guarantees *at least* a locally optimal solution. We further impose the non-negativity and sparsity

constraints [64] as defined below:

$$\begin{aligned} \text{Minimize, } \; E = & \sum_i \Big(Z_s[i,k] - \sum_m \big(X_s[m,k]\, H_s[i-m,k] \big) \Big)^2 \\ & + \lambda \sum_i X_s[i,k]^p \\ \text{Where } X_s[n,k] \geq 0, \; & H_s[n,k] \geq 0, \; \sum_n H_s[n,k] = 1 \end{aligned} \tag{6.2}$$

where we also constrain the $H_s[n,k]$ to sum to 1 for each k, this avoids scaling problems. Note that sparsity implies that while a small number of spectral components in X_s are expected to exhibit high values, most other components have very small (negligible) values. Note also that of the many ways that exist to include sparsity constraints in an NMF framework, we choose to use the L_1-norm. The first term in the objective function (6.2) minimizes the mean-squared error and the second term imposes sparsity on X_s. The optimization is solved subject to the stated non-negativity constraints on X_s and H_s. Corresponding to the L_1 norm, we choose $p = 1$ in (6.2).

6.1.1 Minimization of the Objective Function in an NMF Framework

We minimize the objective function in (6.2) by a variant of the gradient descent approach that ensures that the spectral components at the end of each iteration of the gradient descent are non-negative. Noting that for $p = 1$ in (6.2), the derivative of the objective function with respect to X_s is

$$\frac{\partial E}{\partial X_s[n,k]} = -2 \sum_i \big(Z_s[i,k] - Y_s[i,k] \big)\; H_s[i-n,k] + \lambda \tag{6.3}$$

with the X_s update equation being

$$\bar{X_s}[n,k] = X_s[n,k] - \eta_s \frac{\partial E}{\partial X_s[n,k]}$$

where η_s is the learning-rate parameter. Note that in general there is no guarantee that the updated $\bar{X_s}$ is non-negative. However, we can select a special value of η_s to impose non-negativity. We choose

$$\eta_s = \frac{X_s[n,k]}{2 \sum_i Y_s[i,k] H_s[i-n,k] + \lambda}$$

Incorporating the above value of η_s in (6.3), the updates become:

$$\bar{X}_s[n,k] \leftarrow X_s[n,k].\frac{\sum_i Z_s[i,k]H_s[i-n,k]}{\sum_i Y_s[i,k]H_s[i-n,k]+\lambda/2} \tag{6.4}$$

The updates for H_s can be derived in parallel to the X_s updates in (6.4).

$$\begin{aligned}
\frac{\partial E}{\partial H_s[n,k]} &= -2\sum_i \left(Z_s[i,k]-Y_s[i,k]\right) X_s[i-n,k] \\
\bar{H}_s[n,k] &= H_s[n,k]-\eta_s\frac{\partial E}{\partial H_s[n,k]} \\
\eta_h &= \frac{H_s[n,k]}{2\sum_i Y_s[i,k]X_s[i-n,k]} \\
\bar{H}_s[n,k] &\leftarrow H_s[n,k].\frac{\sum_i Z_s[i,k]X_s[i-n,k]}{\sum_i Y_s[i,k]X_s[i-n,k]}
\end{aligned} \tag{6.5}$$

The iterative update is done for a specified number of iterations. Further, given a non-negative initialization, the updates are guaranteed to be non-negative. Eq. (6.4)(6.5) provides iterative updates for the output of a particular sub-band, indexed by k. Similar processing will also be applied individually to each of the sub-bands. The NMF optimization will at least reach a locally optimal solution. While the estimated X_s may not be identically equal to the actual clean spectra, it is expected that the processing will result in a solution for X_s that will be largely dereverberated.

Fig. 6.2 presents both, the general procedure for frequency-domain NMF processing for dereverberation [64] as well as our specific approach using Gammatone spectra. In both cases, the speech signal is first pre-emphasized (PE) with a causal filter having a single zero at $z = 0.97$. It is then windowed and FFT analysis is performed on the windowed signal. In Fig. 6.2(a) which represents NMF processing in the Fourier frequency domain, the NMF decomposition is directly applied individually to each of the FFT channels. In contrast, in Fig. 6.2(b) which represents our method, NMF processing is applied to each individual channel of the Gammatone filtered spectra, and this is followed by an inverse transformation. Gammatone sub-bands are obtained from the Fourier frequency sub-bands via the Gammatone matrix $G_s[k,k\prime]$, that stores Gammatone frequency response for the $k\prime$ Gammatone sub-band against each of the k Fourier frequency sub-bands. NMF processing is applied to the product $Y_s[n,k]G_s[k,k\prime]$. We discuss the key advantages with Gammatone-based processing Sec. 6.2.2. The NMF processed spectra in the Gammatone domain is multiplied by

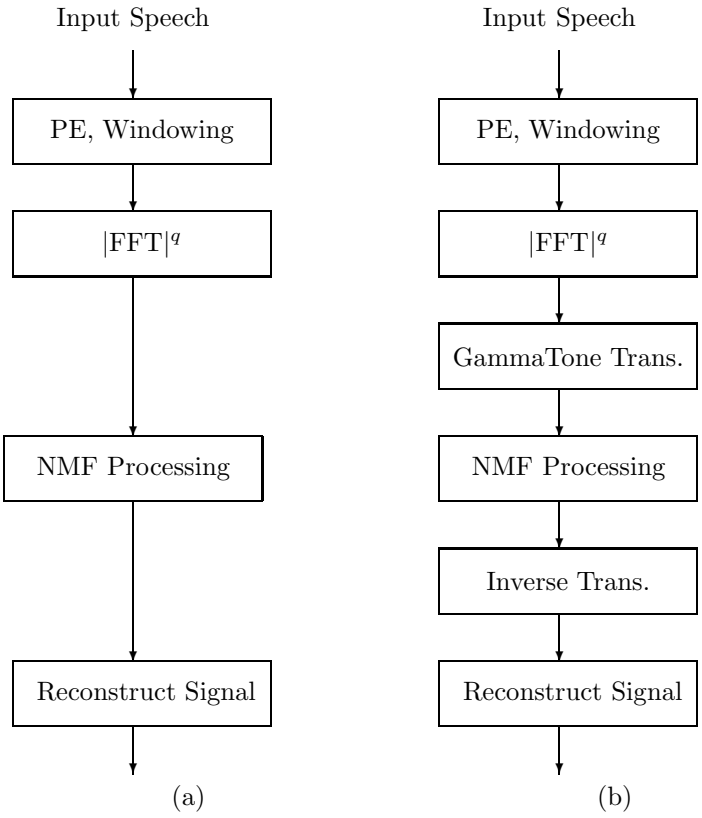


Fig. 6.2: *(a) NMF processing in frequency domain, (b) NMF processing in Gammatone frequency domain.*

the pseudo-inverse of G_s to obtain the processed Fourier frequency components, from which the signal is reconstructed. Since our processing is performed on individual channels in the Gammatone filtered magnitude spectral domain, we call our approach *Gammatone sub-band magnitude-domain dereverberation*. Finally in both the cases the signal is optionally reconstructed or feature vectors for speech recognition may be derived from the resultant dereverberated spectra. For an extension of this work see [74], there we build on the NMF work presented in this chapter to provide an alternate non-negative spectral factorization.

6.2 KEY FEATURES OF GAMMATONE SUB-BAND NMF

In this section we highlight some key aspects of our proposed approach as shown in Fig. 6.2(b).

6.2.1 Advantage of using Magnitude spectra over Power spectra

The model in (6.1) is an approximation and will in general incur an approximation error E_s as follows:

$$Y_s[n,k] = \hat{Y}_s[n,k] + E_s[n,k] = X_s[n,k] * H_s[n,k] + E_s[n,k] \tag{6.6}$$

We have empirically observed that the approximation error E_s is lower in the magnitude spectral domain than in the power spectral domain (see Fig. 3.5). Thus, working in the magnitude-spectral domain incurs lower approximation error. In our approach shown in Fig. 6.2(b) setting the parameter $q = 1$ results in magnitude domain NMF processing, and $q = 2$ results in power domain processing. We will refer to magnitude domain processing as "M-NMF" and power domain processing as "P-NMF".

6.2.2 Advantage of using Gammatone Sub-bands

Processing in the Gammatone domain provides two key benefits. First, the Gammatone sub-bands apply a perceptual weighting to the signal and emphasize the frequency regions where the speech signal is supposed to be dominant for better perception. This directly benefits the quality of the clean signal obtained through the decomposition. Second, working in Gammatone sub-bands offers significant saving in computation. There are about 257 sub-bands for a 512-points FFT in Fourier frequency NMF as compared to about 40 to 80 Gammatone sub-bands for the same processing. We also compared the modeling error in our spectral-domain convolution in Fig. 6.1 (see chapter 3 for details) and found the error to be smaller in the sub-band domain than in Fourier domain. We also obtain a significant practical advantage since we estimate fewer parameters from the same overall data. We refer to Gammatone based NMF processing as "GNMF". We also performed an experiment by substituting the Gammatone-filter spectra with Mel-filter spectra and found the ASR results to be comparable. This suggests that Gammatone spectra may not be uniquely the best for

sub-band NMF processing. Mel filters however have zeroes in their Mel-spectral values. This leads to stability issues in the pseudo-inverse calculation stage, and consequently we prefer Gammatone filters for reliable solutions.

6.2.3 Using Different H_s for Different Sub-bands

In general, we expect the $H_s[n,k]$ in (6.1) to be different for each of the different sub-bands indexed by k. This is expected to result in a more effective solution for X_s than can be obtained by using the same H_s for all the sub-bands. We verify this empirically in the experimental section. To use the same H_s across all sub-bands, the updates in (6.4) can be adapted as follows:

$$\bar{H}_s[n,.] \leftarrow H_s[n,.].\frac{\sum_k \sum_i Z_s[i,k]X_s[i-n,k]}{\sum_k \sum_i Y_s[i,k]X_s[i-n,k]} \tag{6.7}$$

We refer to NMF with the same H_s across all sub-bands as "NMF-H".

6.3 EXPERIMENTAL RESULTS

We applied the NMF formulation in Sec. 6.1 to the problem of dereverberation for ASR. See Sec. 4.3.1 for our ASR experimental setup. In our experiments, we simulated reverberation effects to various degrees in the the DARPA Resource Management (RM) database, dereverberated the signals, and then measured the recognition accuracy on dereverberated signals using matched and mismatched recognizers.

Utterances in the RM database were artificially reverberated with different RTs [75], as shown in Fig. 7.12. In the first experiment, NMF processing methods as shown in Fig. 6.2 were applied to dereverberate the utterances. We used 15-20 iterations of NMF processing with a window size of 64 ms for the NMF processing, reconstructed the speech, and extracted conventional MFC features for ASR from the reconstructed speech. These use a window size of 25 ms. In Fig. 6.3, we present our experimental results with the different implementations of the NMF processing described in Secs. 6.2.1, 6.2.2, and 6.2.3. Note that the bar entitled "P-NMF" shows ASR results for conventional sparsity constrained power domain NMF [64], while the bar titled "M-GNMF" shows the performance obtained with our approach. Experimentally, we found that sparsity was not helpful for the Gammatone sub-bands and hence was not

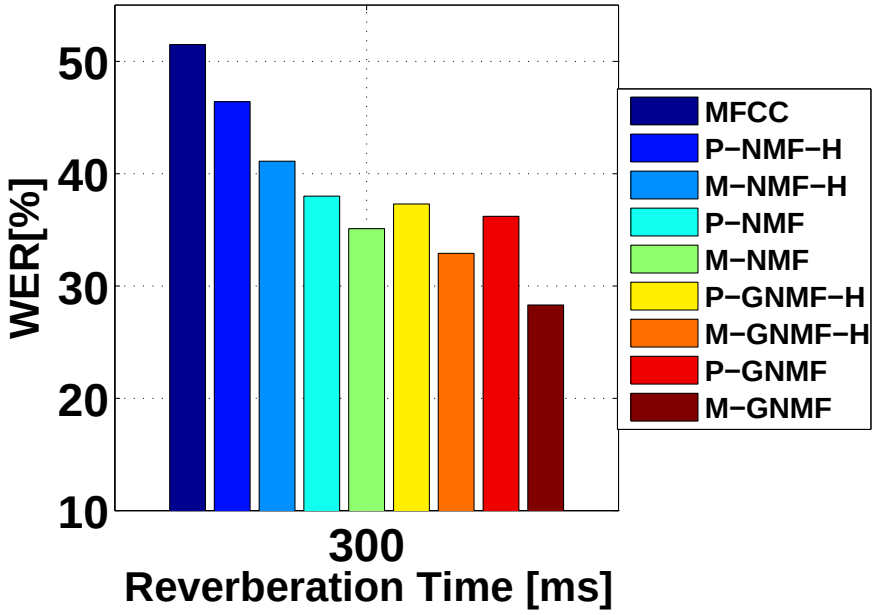


Fig. 6.3: *WER comparisons for different flavors of NMF.*

applied. A small sparsity [64] factor was applied for NMF processing in the Fourier frequency domain.

Overall, we note that NMF processing in gammatone bands provides 20-25% relative reduction in word error rate (WER) over the same processing in the Fourier frequency domain. NMF processing in the magnitude domain provides 10-18% relative improvement over power domain processing. In addition, the use of different H_s for different sub-bands provides a 10-20% relative improvement over the same H_s for all sub-bands. We thus finalize NMF processing in the magnitude domain with Gammatone sub-bands and different H_s for different sub-bands as our baseline NMF processing, and from now on we refer to the configuration "M-GNMF" as simply "NMF".

In Fig. 6.4 we plot the WER obtained from the Fourier-domain-based magnitude NMF processing. We study the effect of sparsity in clean as well as in RT of 300 ms condition. We see that an increasing value of sparsity provides a tradeoff between the WER for clean and the reverberated conditions. Though sparsity helps in reverberant conditions, a large sparsity value increasingly hurts results in the clean condition.

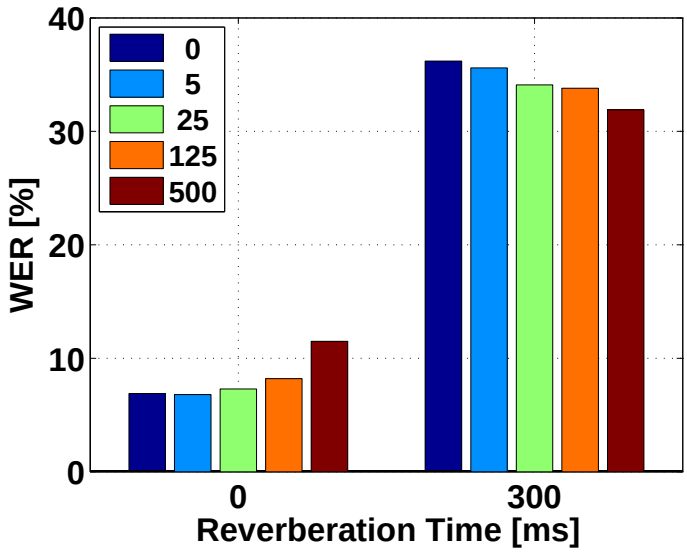


Fig. 6.4: *WER comparisons for m-NMF in Fig. 6.3 with different sparsity factors.*

Since the spectra of clean speech are already sparse, we need not enforce additional sparsity constraints on the clean condition. Based on the tradeoff in Fig. 6.4, we choose a sparsity level of 5 for the Fourier-domain-based NMF processing. The Gammatone speech spectra do not exhibit sparsity characteristics and hence we did not incorporate a sparsity factor in GNMF conditions in Fig. 6.3.

In Fig. 6.5, we plot the WER results for the case where the system is trained with clean (unreverberated speech) and tested on dereverberated speech. We see that here the relative reduction in WER is limited to 15-20% for the baseline dereverberation algorithms. The NMF processing provides a 45% relative reduction in WER at RT of 300 ms which is substantially better than any of the baseline algorithms. We also note that improvements from NMF processing are similar to those obtained from LIFE processing in Chapter 5. A key advantage of NMF is that it does not require any *a priori* information about the distribution of speech features. Although, LIFE processing in general requires such prior distribution information, we experimentally evaluated in Sec. 5.3 that LIFE processing with single Gaussian density worked the best. Correspondingly, LIFE processing too does not require *a priori* density information.

In Fig. 6.6, we plot the unprocessed and NMF compensated speech spectrograms. The effect of reverberation can be seen in the lateral spectral smearing in unprocessed

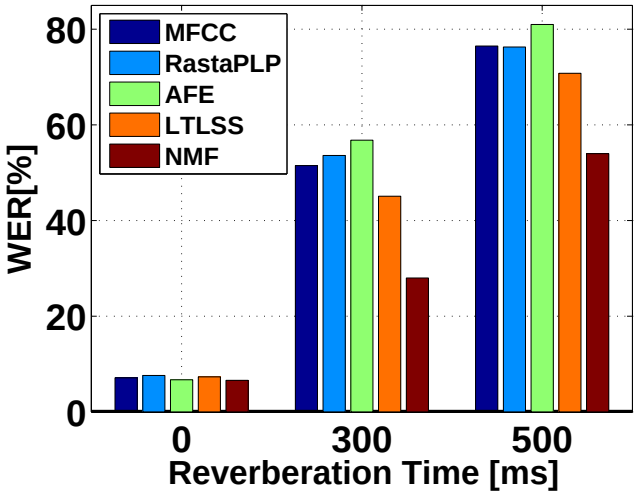


Fig. 6.5: *NMF WER comparisons for clean-training.*

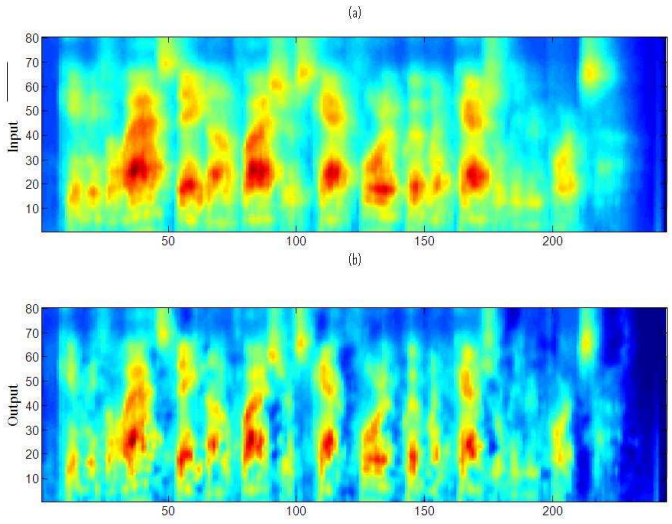


Fig. 6.6: *[Top] Unprocesssed Spectra, [Bottom] NMF Processed Spectra.*

spectr. Reverberation not only blurs the boundaries between different word units but also makes them less distinct from one another. The NMF-processed spectrogram is also shown in Fig. 6.6. There, we see that the smearing along time is reduced, the word boundaries are clearer and the amount of overlap from one word segment to another is also reduced.

6.3.1 An Oracle Experiment with Zero Modeling Error

In Sec. 5.3.4 we noted that the reverberation model on which LIFE processing operates is imperfect and carried out an oracle experiment to study LIFE processing under zero modeling error. NMF processing works on the reverberation model in (3.10) that too is affected by an additive error term. In this section, we report an NMF-based oracle experiment that parallels the LIFE-based oracle experiment in Sec. 5.3.4. This isolates the performance of NMF by itself under zero modeling error conditions. We simulated the oracle experiment by directly convolving the spectral features with an exponentially decaying RIRs corresponding to RTs of 300 and 500 ms. NMF processing was applied on the convolved features and was blind to the actual RIR. We present the corresponding ASR results in Fig. 6.7. We see that NMF-based processing with zero modeling error achieves a WER of 21% at RT of 500 ms. In Fig. 6.5 we reported a comparable practical (non-oracle) NMF experiment and noted that results using oracle knowledge were substantially better. Thus we believe that modeling error substantially limits the NMF performance in practical environments. Comparing the oracle experiments for NMF and LIFE processing in Fig. 6.7, we see that LIFE processing shows greater potential than NMF.

6.3.2 Experiments using Real Room Impulse Responses

In the experiments so far, the reverberated speech was obtained by convolving clean speech with a simulated room impulse response (RIR). In this section, we apply NMF processing on data obtained from a real RIR. We obtained an RIR corresponding to RT of 500 ms from the ATR [13] database. In Fig. 6.8, we plot the WER for MFCC and NMF processing. We see that benefits from NMF processing extends to real RIRs, and at RT of 500 ms, we obtain 38% relative reduction in WER.

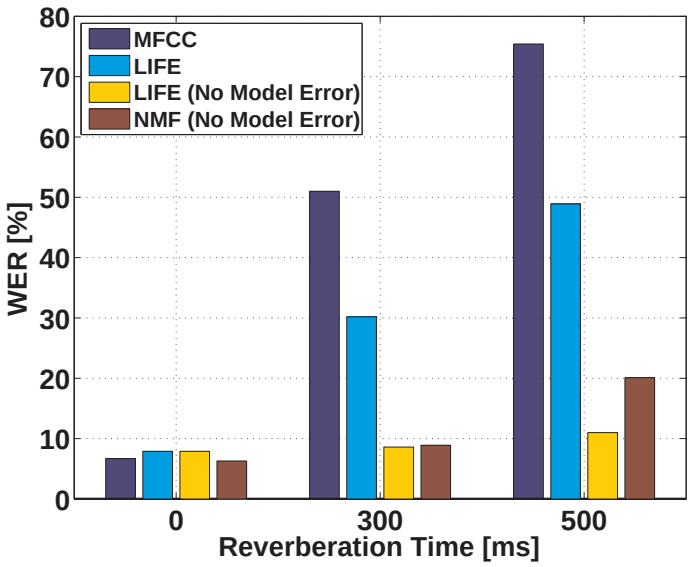


Fig. 6.7: *NMF processing for an oracle experiment with zero modeling error.*

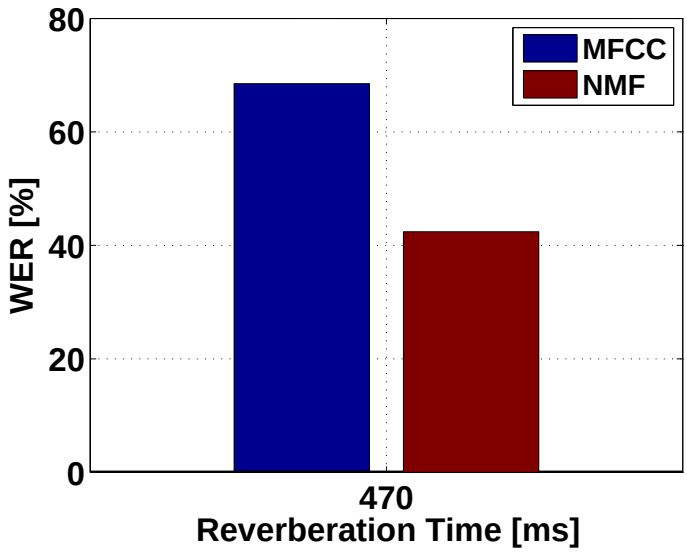


Fig. 6.8: *WER comparisons for NMF processing with a recorded RIR.*

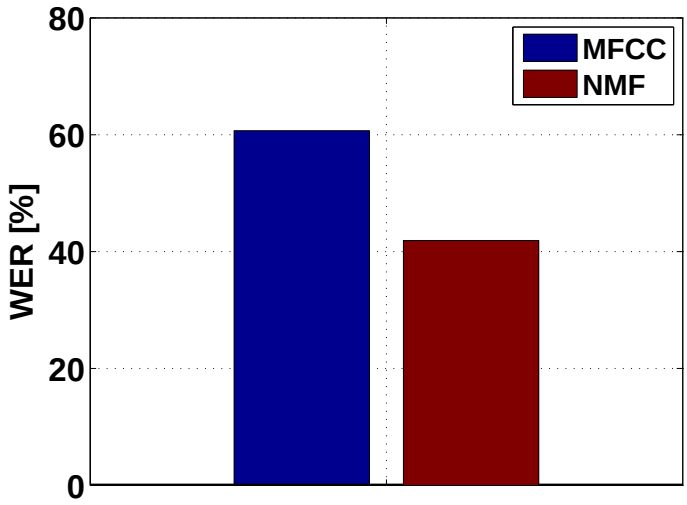


Fig. 6.9: *WER comparisons for NMF processing on recorded speech.*

6.3.3 Experiments on Recorded Speech

In the experiments so far, the reverberated speech was obtained by convolving clean speech with either a simulated RIR or a real RIR. In this section, we apply NMF processing on speech recorded as part of ATR [13] database. The database consists of segments from the TIMIT database. In Fig. 6.9, we plot the WER for MFCC and NMF processing. We see that benefits from NMF processing extend to recorded speech where, we obtain 31% relative reduction in WER. The recorded ATR database inadvertently included additive noise with SNRs in the range of 15-20 dB. We believe that NMF proceeding could have provided greater improvements in absence of the additive noise.

6.3.4 Multi-style Experiments

In all the results reported so far in this chapter, the experimental paradigm was clean training. In this section, we introduce a multi-style paradigm for speech recognition. In multi-style, the training set includes data from multiple available environment conditions. For additional details on the multi-style training paradigm, see Sec. 5.4.4.

In Fig. 6.10, we plot WER for the case when the ASR system is trained on the same kind of speech as it is tested on (matched-condition training), specifically the

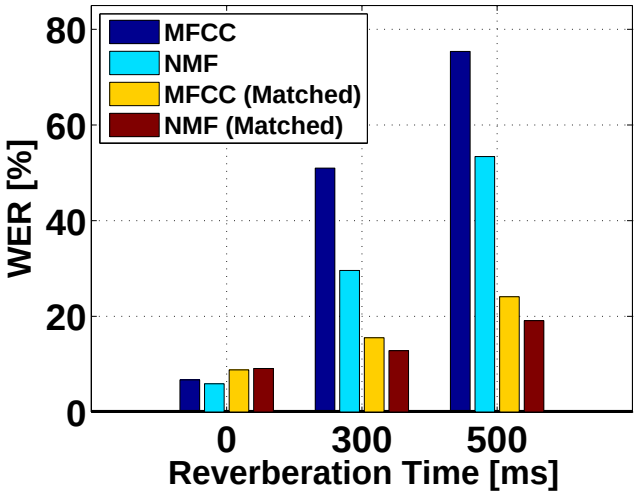


Fig. 6.10: *NMF WER comparisons for matched-training.*

training set consists of data from RTs of 0, 300 and 500 ms. We note from those experiments that the algorithm is able to improve over matched-condition testing, and to substantially improve over clean-condition testing. Note that it is usually very difficult to improve over matched-condition testing in ASR. It is usually used as a gold-standard in many instances. Matched NMF provides an additional 19% relative reduction in WER over simple matched training and testing with MFCs from reverberated speech.

6.4 DISCUSSION

In this chapter we presented an NMF-based approach for dereverberation of speech signals in the Gammatone sub-band domain. This work has specific advantages for ASR that we have experimentally shown to be valid. The algorithm results in 30-45% WER reduction under mismatched conditions, when the system is trained on clean speech but attempts to recognize reverberated speech, or dereverberated speech as the case may be. The algorithm presented is able to improve WER performance by about 15% relative to matched-condition training, which has been generally observed to be a performance threshold that is very hard to exceed.

CHAPTER 7

DELTA SPECTRAL FEATURES FOR ROBUST SPEECH RECOGNITION

In Chapters 4, 5, and 6, we studied the problem of dereverberation for ASR and provided algorithms to compensate for it. Apart from reverberation, additive noise is also a major challenge for ASR. Further, many of the ASR usage environments may include both noise and reverberation [18]. Because ASR word accuracy degrades [25, 76, 77] severely in additive noise conditions, we consider the issue of robustness of ASR to additive noise in this chapter.

Even though a number of algorithms [53, 77–80] have been successfully developed for noise robustness of ASR, the improvement is still limited in many of the real-world noise conditions. A few algorithms have shown very good performance in certain noise conditions, but the performance could not be replicated across different noise conditions. Our work in this chapter is guided towards deriving inherently robust features for ASR that can provide improvements across different noise conditions. We base our work on some of the key characteristics of the speech signal that differentiate it from noisy signals, specifically the non-stationarity characteristics of speech.

Most current speech recognizers derive their features in the broad framework as shown in the left column of Fig. 7.1, which describes the development of features similar to Mel-frequency cepstral coefficients (MFCC). Typically delta-cepstral and double-delta cepstral coefficients are appended to MFCC features, as discussed below.

In this chapter we argue that recognition accuracy in many practical environments is improved by replacing delta features in the cepstral domain by delta features in the *spectral* domain. We support this argument using both graphical and analytical arguments based on the spectra of speech and common environmental noises, as well as experimental studies in which we compare the recognition accuracy obtained using our framework in the recently-proposed robust ETSI advanced front end (EAFE) [31] and power-normalized cepstral coefficients (PNCC) [81].

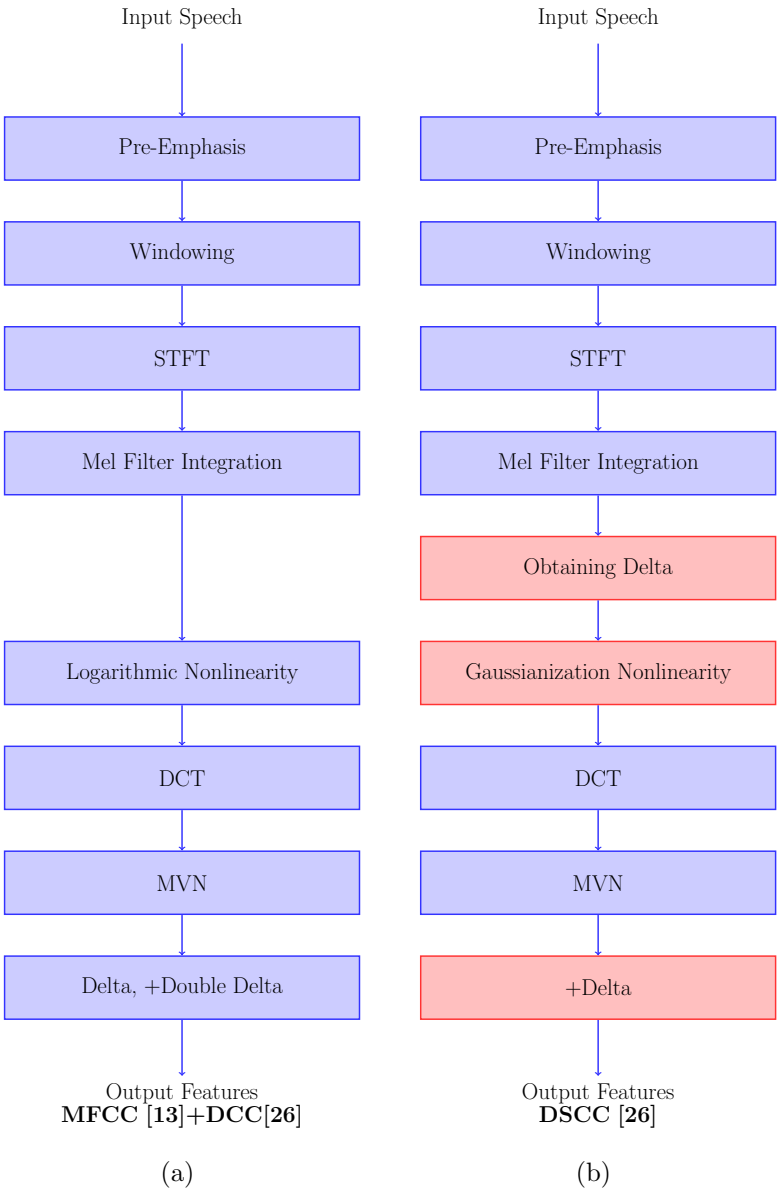


Fig. 7.1: *(a) 13-dimensional MFCC features and 26-dimensional delta-cepstral coefficients (DCC), (b) 26-dimensional delta-spectral cepstral coefficients (DSCC) features.*

The rest of the chapter is organized as follows: we discuss the delta-cepstral features and their robustness to noise in Sec. 7.1. In Sec. 7.3 we propose the new delta-spectral features. We provide the rationale for our proposed features in Sec. 7.4, and

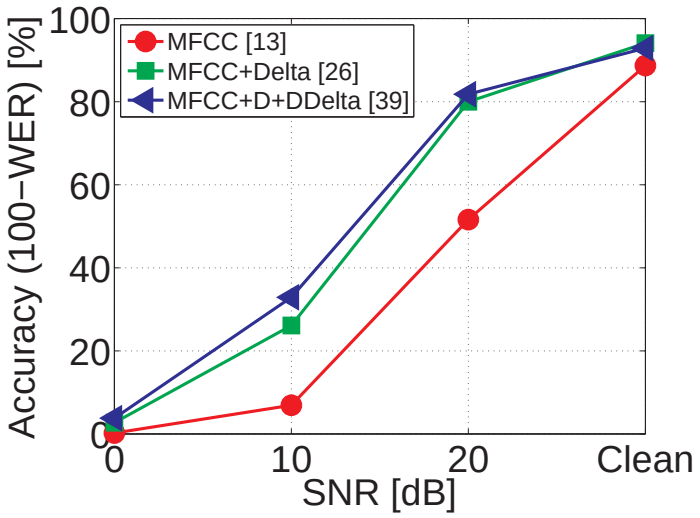


Fig. 7.2: *Word error rates (WERs) obtained in additive white-noise using MFCC features, MFCC+Delta features, and MFCC+Delta+DoubleDelta features.*

our experimental results are in Sec. 7.5. Sec. 7.6 summarizes this study.

7.1 DELTA-CEPSTRAL FEATURES

Delta-cepstral features were proposed (in a different form) by Furui in [82] to add dynamic information to the static cepstral features. They also improve recognition accuracy by adding a characterization of temporal dependencies to the frames of the hidden-markov models (HMM), which are nominally assumed to be statistically independent of one another. For a short-time cepstral sequence $C[n]$, the delta-cepstral features are typically defined as

$$D[n] = C[n+m] - C[n-m] \tag{7.1}$$

where n is the index of the analysis frames and in practice m is typically 2 or 3. Similarly, double-delta cepstral features are defined in terms of a subsequent delta-operation on the delta-cepstral features. Fig. 7.2 plots the word error rate (WER) for speech recognition in the presence of white noise for the DARPA Resource Management (RM) database, following experimental procedures described in Sec. 7.5. We note that the addition of delta-cepstral features to the static 13-dimensional MFCC features

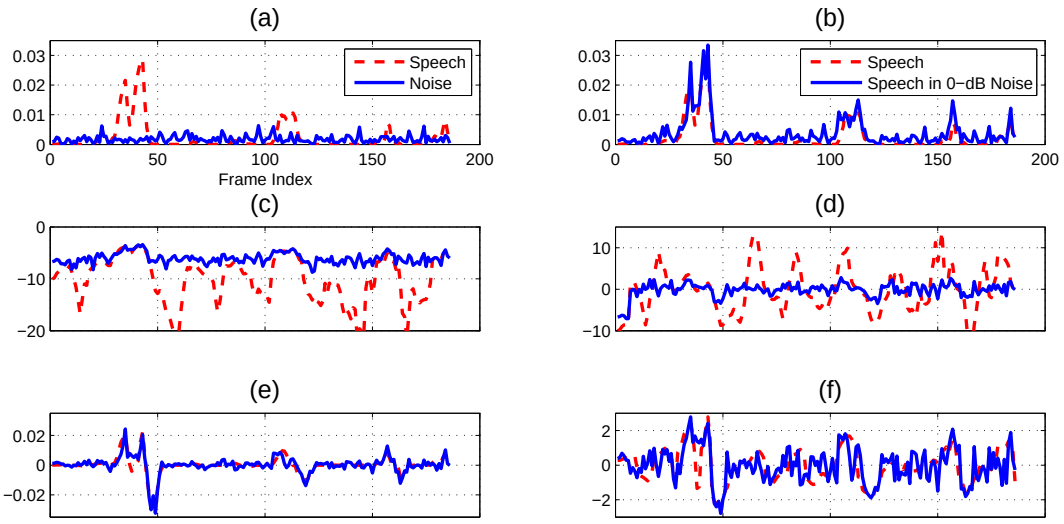


Fig. 7.3: *(a) Short-time power plot of a Mel-channel (center frequency 1000 Hz) for a speech and a "real-world" noise segment using 10-ms frames, (b) Short-time power for clean speech as in (a) and speech in 0-dB "real-world" noise from (a), (c) Logarithmic power plot for clean speech and noisy speech in (b), (d) Temporal difference operation over the signals in (c), (e) Temporal difference over the signals in (b), (f) Gaussianization operation over the signals in (e).*

greatly improves speech recognition accuracy, and a further (smaller) improvement is provided by the addition of the double-delta cepstra. For these reasons some form of delta and double-delta cepstral features are part of nearly all speech recognition systems. It can be seen that the improvement provided by delta features gradually diminishes with lower SNR. We also note that from Eq. (7.1), it can be easily shown that $E[D[n]C[n]] = 0$, where $E[.]$ is expectation operator. So the delta-features are uncorrelated with the static features which helps the frame-independence assumption in the HMM in ASR.

While the addition of delta-cepstral coefficients (DCC) to MFCC coefficients does indeed improve ASR recognition accuracy, they do not provide great robustness in low

SNR noise conditions. The reasons for this can be understood in graphical form by consideration of Fig. 7.3, which depicts various manipulations of the short-time power of clean speech, and speech in "real-world" noise at 0-dB SNR (with noise recorded naturally from locations such as a market, a food court, the street, and a bus stop). Fig. 7.3(a) plots the short-time power for a particular speech segment, and for the corresponding noise segment. We note that the speech signal power exhibits a very high dynamic range, while the noise spectral power is much more static than the speech power. Fig. 7.3(b) plots the short-time power for clean speech and speech plus noise at 0 dB noise using the noise from Fig. 7.3(a). Unsurprisingly, the peaks of Fig. 7.3(b) remain relatively intact, while the "valleys" are filled by the noise. The corresponding log-power values are shown in Fig. 7.3(c), and they are a step in the extraction of MFCC coefficients, as seen in Fig. 7.1(a). Due to compressive nature of the log nonlinearity, the spectral peaks are approximately same for the clean and noisy speech but the remaining frames exhibit a high degree of mismatch. Since noise fills the valleys of the curves, it is relatively stationary, the noisy log-spectral contour exhibits a sharply reduced dynamic range in comparison to the corresponding clean log-spectral contour. Finally, plotting the corresponding delta-cepstral features in Fig. 7.3(d), we note that the delta features still exhibit a high degree of mismatch between clean and noisy conditions. The delta-spectral features proposed in the Sec. 7.3 both retain the contextual properties of delta-cepstral features and are robust to noise and reverberation as well.

7.2 NON-STATIONARITY OF SPEECH POWER SEQUENCE

It is well known that speech signals are locally stationary within 25-30 ms but globally non-stationary. In addition, speech signals exhibit a very high dynamic range in terms of their spectral amplitude in a narrow band frequency range. In Fig. 7.4, we plot the output of a Mel-filter bank centered at 1050 Hz for a typical speech and noise signal. The noise signal is part of a real-world noise recording (see details in Sec. 7.5). We observe that the noise power flow is relatively more stationary than speech power flow. We also observe that the speech power flow exhibits very high amplitude variations. The stationarity and dynamic range characteristics of the speech power flow is very particular to speech signals – it is not exhibited by many other real-world noise conditions. It is probably these characteristics of speech signals that provide human ears

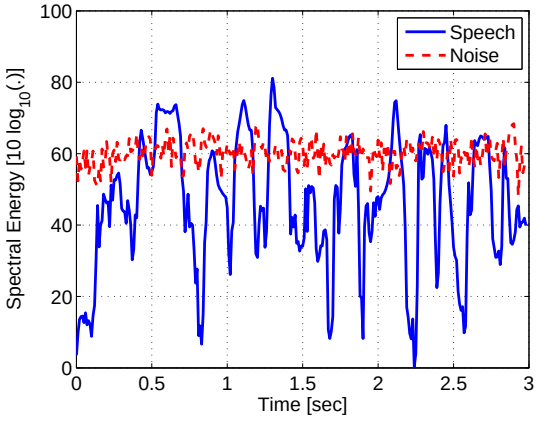


Fig. 7.4: *Output of* 14^{th} *Mel-frequency filter, center frequency 1050 Hz, for a typical speech and noise segment.*

with the enormous ability to detect speech sounds in presence of background noise [83]. Due to the differences in stationarity characteristics of speech and noisy signals, human ears can largely ignore noise that appears to be perpetually present [84] and pay more attention to the speech signal with rapidly changing spectral patterns. Among the different types of interfering signals that we face in the real-world, interference from another speech signal is among the most difficult speech perception tasks for humans. This is because the interfering speech signal will, of course, have similar power flow characteristics as that of the target speech signal, so the ability to distinguish one speech signal from another diminishes.

Based on our understanding and analysis on the differences of speech and noise spectral characteristics, we propose delta-spectral cepstral coefficients (DSCC) features for robustness to noise. As shown in Fig. 7.1, noise spectral values are relatively flat whereas speech spectral values change rapidly. So taking a difference across frames strongly attenuates the noise components.

7.3 DELTA-SPECTRAL CEPSTRAL COEFFICIENTS

We now discuss the delta-spectral cepstral coefficients for ASR. These features are motivated by the non-stationarity of speech signals that had been observed in Sec. 7.2 and also in Fig. 7.3(a) where it is easily observed in that figure that the short-time

power of speech varies much more rapidly than the short-time power of noise. The vast differences between the rate of change of power for speech and noise are likely to be one of the many cues that human ears can use to ignore the relatively stationary noise signals and focus on the rapidly-changing power of speech signals.

The proposed delta-spectral ceptral coefficient (DSCC) features are described in block diagram form in Fig. 7.1(b). Our objective is to combine the speech contextual information captured by the DCC features in Fig. 7.1(a) with a greater degree of robustness to additive noise. As can be seen, the major changes are that the initial time-differencing operation is now moved earlier in the processing and a new Gaussianization stage is added. Specifically, performing the delta operation described by Eq. (7.1) in the spectral domain will enhance the fast changing speech components, and suppress the slowly-changing noisy components. Fig. 7.3(e) plots the outcome of the delta operation in the spectral domain on the power contours in Fig. 7.3(b). The advantage of the delta-spectral approach is clear by comparison of the similarity of the curves representing clean and noisy speech in Fig. 7.3(e) (which were obtained by applying the delta operation in the spectral domain) to the corresponding curves in Fig. 7.3(d) (which were obtained by applying the delta operation in the cepstral domain). Thus, overall the DSCC feature extraction process is identical to that of MFCC features until the "Mel-Filter Spectrum". The DSCC features are derived from the Mel-spectrum by applying a temporal difference operation on the spectral values. Specifically, a filter

$$H_d(z) = z^d - z^{-d} \tag{7.2}$$

is applied on the individual Mel-spectral feature sequences. The parameter $d \in [2, 4]$ was experimentally found to work well.

Experimentally we found that the delta-spectral features in their current form are unsuitable for speech recognition applications because the raw delta-spectral cepstral features are highly non-Gaussian, as is seen in Fig. 7.5. To adapt the delta-spectral features for speech recognition, we apply histogram normalization to the delta-spectral features to give them a Gaussian distribution, as shown in Fig. 7.5(b). This Gaussianization nonlinearity is applied on an utterance-by-utterance basis. Fig. 7.3(f) plots the "Gaussianized" delta-spectral features, which are numerically compressed by the DCT operation as in Fig. 7.1(b). The DCT operation compresses the 40-dimensional delta-spectral features to a 13-dimensional vector of delta-spectral cepstral coefficients

(DSCC). Double-delta features are then derived from the delta-spectral features in the cepstral domain.

7.3.1 DSCC Comparison with MFCC Features

There are several key differences between the DSCC and MFCC features. The DSCC features completely ignore the static-spectral contents, deriving their features instead entirely from the dynamic-spectral contents. Tracing the evolution of the ASR, we find that the static features were the first to be applied for ASR, dynamic features were later appended to the static features to further improve ASR performance. But as we will verify in Sec. 7.5, the dynamic features in the DSCC features are not only good for ASR but they are also very robust to additive noise. Another important difference between the DSCC and MFCC features is a different nonlinearity in the DSCC features. The log nonlinearity has been motivated in the ASR community from human auditory models that incorporate a nonlinearity stage. In DSCC we instead apply a data-driven "Gaussianization" nonlinearity.

7.3.2 Comparison of DSCC and RASTA Features

A key difference between the DSCC features and RASTA [25] processing is that the filtering operation in RASTA is strongly linked to the modulation spectral characteristics where it attenuates the modulation frequencies outside the range of 4-20 Hz. The DSCC features are not based on the modulation characteristics and are instead based on capturing the dynamic transition characteristics in speech. Another significant difference between the two is that the filtering operation in DSCC is applied on the linear spectral values, whereas RASTA is applied in a compressive nonlinearity domain. Since noise is additive in the spectral domain if the sources are independent, the filtering operation in DSCC strongly suppresses noise, whereas RASTA processing can only partially suppress noise.

7.3.3 DSCC Comparison with DPS Features

The DSCC features can be seen as a continuation of DPS. The work in [85] noted the additivity of noise in the linear spectral domain and proposed a difference operation

directly on the spectral values. Our work differs from DPS in the subsequent processing of the delta-spectral values. Our work as well as the work DPS faced the problem of handling the resulting negative-spectra from the difference operation. DPS handled the negative spectra by taking absolute values. This omits the distinction between the rising and falling spectral regions, which in fact become indistinguishable in the absolute values and harm the robustness of ASR. DSCC features preserve the aforesaid distinction and apply a Gaussianization nonlinearity, the DSCC ASR results show substantial improvement over DPS.

7.4 DSCC FEATURE ANALYSIS

In this section we provide a more formal analysis of the SNR improvement in white noise using the DSCC features. Assuming that the noise is a white Gaussian sequence sample distribution w_i of the form $\mathcal{N}(0, \sigma^2)$, the power P in an independently-observed set of N samples is $P = \frac{1}{N}\sum_{i=1}^{N} w_i^2$. P follows a chi-square distribution with N degrees of freedom (DOF), which becomes approximately Gaussian for large N. Under the Gaussian assumption for P, it can be shown that

$$E[P] = \frac{1}{N}E[\sum_{i=1}^{N} w_i^2] = \sigma^2$$

$$\begin{aligned} Var[P] &= E[P^2] - E[P]^2 = \frac{E\big[\sum_{i,j} w_i^2 w_j^2\big]}{N^2} - \sigma^4 \\ &= \frac{1}{N^2}\Big(\sum_i E[w_i^4] + \sum_{i,j,i\neq j} E[w_i^2 w_i^2]\Big) - \sigma^4 = \frac{2\sigma^4}{N} \end{aligned}$$

Thus, P is approximately distributed as $N(\sigma^2, \frac{2\sigma^4}{N})$. The DC power associated with P is the square of the mean, σ^4, while the AC power is the variance $\frac{2\sigma^4}{N}$. DSCC processing removes the DC power, and we can express the impact of this effect using the ratio

$$\text{Noise suppression} \approx -10\log_{10}\Big(\frac{Pow_{AC}}{Pow_{AC} + Pow_{DC}}\Big) \tag{7.3}$$

$$= 10\log_{10}\big(1 + N/2\big) \tag{7.4}$$

We use a speech analysis window duration of 25 ms, so the number of samples in the window duration becomes $N = 400$ with a sampling frequency of 16,000 Hz, and for

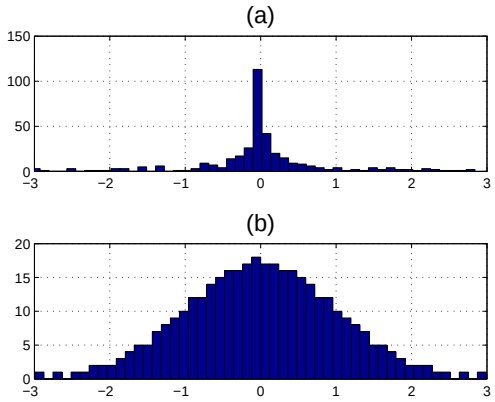


Fig. 7.5: *Histogram of short-time power after the delta operation for a clean-speech sample (a) before and (b) after Gaussianization.*

$N = 400$, the consequent white noise suppression is 23.03 dB. Thus, the maximum possible benefit with DSCC processing is a 23-dB SNR noise suppression for the white noise case.

Noise Type	White	Real-World	Music
Predicted noise suppression	23	12	3.5
SNR threshold-shift in ASR	8.3	7.5	5

Table 7.1: *Predicted noise suppression and observed SNR threshold-shift in an ASR experiment for different noise conditions (in dB). The prediction and observation exhibit a correlation coefficient of 0.93.*

In Table 7.1, we experimentally derive the degree of noise suppression for different noise conditions based on the percentage of total power that is DC power, as above. As expected, the noise suppression so obtained is greater for relatively stationary noises such as white noise and the "real-noise" conditions than for background. We also present the experimentally-observed shift in effective SNR that will be discussed below in conjunction with a speech-recognition task (*cf.* Fig. 7.11). While the observed SNR shifts are not equal to the calculations above for many reasons, (including suppression of both speech and noise at other frequencies imposed by the DSCC algorithm and

subsequent nonlinearities in processing), the trends of the dependencies are similar. Specifically, we found the correlation coefficient between the prediction and observation in Table 7.1 to be 0.93, that indicates a very good prediction performance for the measure in (7.3).

7.4.1 Empirical Distortion Analysis at Different Feature Stages

In this section we empirically study the different feature sequences in Fig. 7.3 with respect to their robustness to additive noise at the speech-signal level. Fig. 7.3(b) plots the power sequence for a segment of clean speech and speech in 0-dB noise for a particular frequency-band as noted in the figure. As expected the additive noise distorts the original power sequence. This distortion eventually leads to a mismatch between the features derived from clean speech and those from noisy speech, leading to a loss in accuracy for an application that is trained to work on features derived from clean speech. Similarly, additive noise also affects the log-power, delta-log-power, delta-power and Gaussianized-delta-power feature sequences. In this section we empirically quantify distortion at the different feature stages for the DCC and DSCC features and attempt to derive insights into their potential benefit in speech application. For convenience we define the following terms:

1. Signal-to-noise ratio (SNR): We use the conventional definition of SNR, which is the ratio of speech signal power to that of noise signal power (in dB).
2. Power-spectrum-to-distortion ratio (PSDR): The error between the power-spectral sequence corresponding to clean speech and that corresponding to speech in noise is termed as distortion in the power-spectral domain. We quantify it in terms of the power ratio of the clean power-spectral sequence and this distortion signal.
3. Log-power-spectral-to-distortion ratio (LPDR): Similar to PSDR, we quantify the distortion in the log-power-spectral domain in terms of LPDR.
4. Delta-power-spectral-to-distortion ratio (DPDR): DPDR quantifies the distortion in the delta-power-spectral sequences.
5. Gaussianized-delta-power-spectral-to-distortion ratio (GDDR): GDDR quantifies the distortion in feature sequences that subsequently form the DSCC fea-

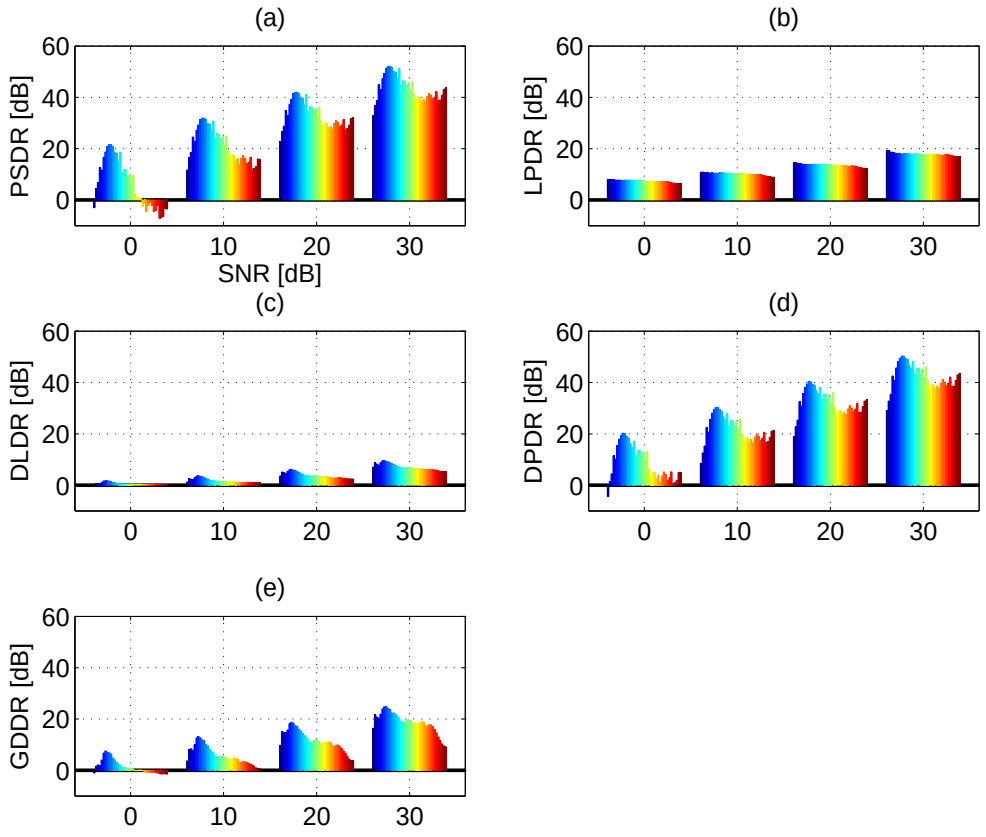


Fig. 7.6: *Feature to distortion power ratio for an additive white noise signal at different input SNR levels, (a) Power-spectral to distortion, (b) Log-power to distortion ration, (c) Delta-log-power to distortion ratio, (d) Delta-power to distortion ratio, (e) Gaussianized-delta-power to distortion ratio.*

tures.

Following the above definitions for SNR and distortion at the different feature levels, we relate them in Fig. 7.6. We considered white noise in the above experiment. The horizontal axis denotes SNR in dB. At each SNR, we plot the distortion levels for each of the 40 Mel-channels, with the channels in ascending numbers. We make the following observations in Fig. 7.6:

- PSDR is approximately linearly related to SNR for high SNR regions. The relationship is however nonlinear in low SNR regions, especially for high frequencies. Further even at 0-dB SNR, the frequencies in 1000 Hz region possess 20-dB PSDR but the high frequencies exhibit negative PSDR levels. Thus as expected the high frequencies are more adversely affected in white noise conditions. Although the above observation has been known for years, a quantification of the observation is new. We see that at 20-dB SNR, the PSDR difference between 1000-Hz region and high-frequency regions is 10 dB but for 0-dB SNR, the difference becomes 30 dB.
- In general at different SNRs, the LPDR levels are significantly smaller than PSDR levels. Thus the conventional log operation on the power-spectral sequences hurts the distortion-robustness aspect of the power-sequences. The log-operation was conventionally applied to mimic the human auditory models that include a compressive nonlinear stage. Although the compressive log-nonlinearity provides a way closer to the auditory models, the quantification in Fig. 7.6(b) highlights that compared to the power sequences, the log compressed power sequences are less robust to additive noise. We note a similar tradeoff between compressive nonlinearity and noise robustness in the DPDR and GDDR levels. Gaussianization is also a compressive nonlinearity and comparing Fig. 7.6(d) and (e), the absolute distortion levels decrease with nonlinearity.
- The DPDR values in Fig. 7.6(d) are similar to that of PSDR, and in fact better for high frequencies in 0-dB SNR regions. This is understandable since additive noise is relatively stationary with respect to the nonstationarity in speech, consequently noise is partially nullified due to delta-operation on the power-sequences.
- Fig. 7.6(e) and Fig. 7.6(c) compare distortion levels that respectively apply to the DSCC and the DCC features. We see that the GDDR levels are significantly higher than the DLDR levels. At 10-dB SNR, the GDDR levels are approximately 10-dB higher than DLDR levels. The DSCC features are thus expected to be more noise robust than DCC features.

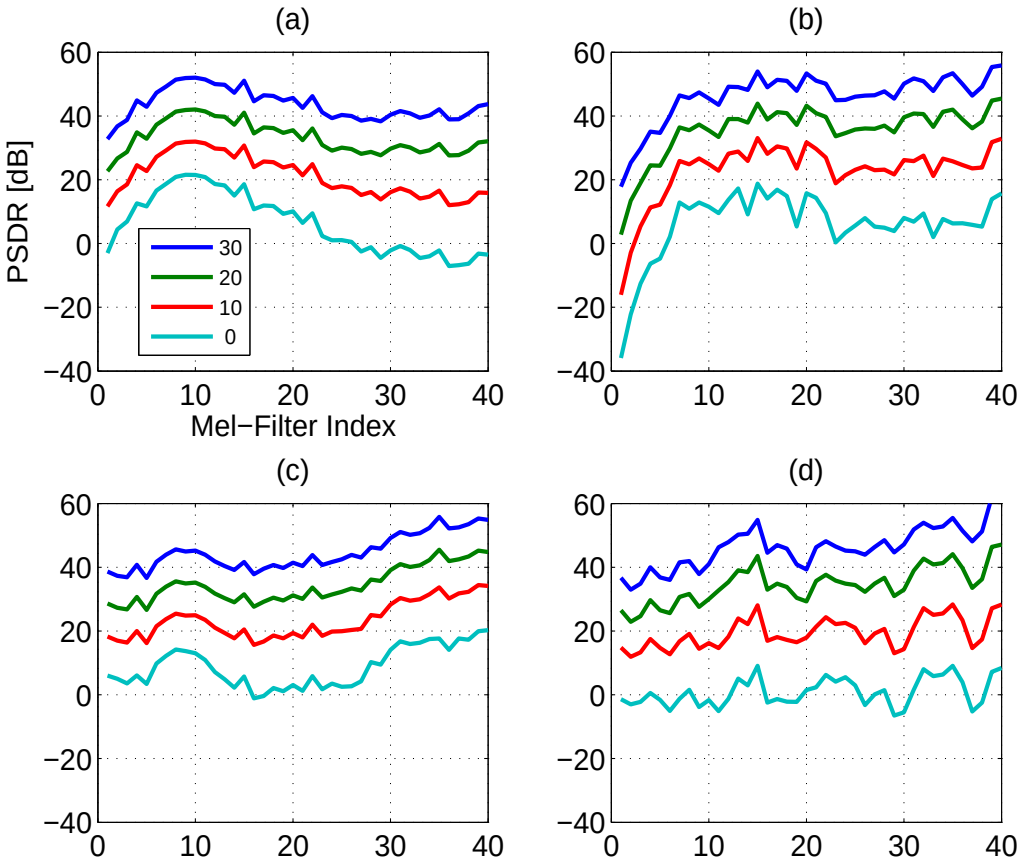


Fig. 7.7: *PSDR for the individual Mel-channels against the SNR levels for additive noises in (a) White-noise, (b) Background music, (c) Real-world noise, (d) Interfering speaker. The legends in the plot indicate the SNR in dB.*

In Fig. 7.6 we related SNR in the speech-signal domain and distortion in the features for white noise condition. Next, we expand our study to different noise types. This new study will highlight the characteristics of different noises and the associated effect on the speech frequency bands. In Fig. 7.7, we plot the PSDR levels for each of the 40 Mel-channels for 4 different noise types at different SNR levels. Fig. 7.7(a) is essentially an alternate representation of the information in Fig. 7.6(a). We see that white-noise increasingly affects the high-frequency regions with the Mel-channel regions corresponding to 1000-Hz region being the most robust. Moving from 30-dB to 0-dB SNR, the high frequency regions loose 45-dB PSDR, the low frequencies loose 38-dB PSDR whereas the frequencies around 1000 Hz loose only about 30-dB PSDR. The above quantitative trend is expected since the speech-signal power exhibits a peak in the 800-2000 Hz regions, whereas the white-noise signal power is uniformly distributed across the entire spectrum.

Fig. 7.7(b) plots the PSDR levels for background music conditions. The music sample was obtained from HUB-4 database. We note that the low-frequency regions are the most affected due to music. The high frequency regions possess higher PSDR for background music than for white-noise. Specifically we note that for high-frequency regions, the PSDR for background music is non-negative at 0 dB SNR, whereas the PSDR is below 0-dB for the corresponding white-noise case. Fig. 7.7(c) plots the PSDR levels for noise samples recorded in real-world settings. We observe that the Mel-channels from 18 to 24 that correspond to the frequency band of 1500-2500 Hz are significantly more degraded than other frequency regions. For the interfering speaker case in Fig. 7.7(d), we note a stark difference with the rest of noise conditions in Fig. 7.7(a)-(c). There, the PSDR levels are nearly 0-dB at 0-dB SNR for all the Mel-channels. This was clearly not so for other noises. We also note that for the 1000-Hz frequency regions, the PSDR levels are the lowest for interfering speaker case. This explains and supports that the interfering speaker case may perform the poorest among different noise conditions.

Fig. 7.8 demonstrates a comparison between the DSCC and DCC features in terms of their respective distortion measures in GDLR and DLDR. We note that DSCC exhibits superior distortion levels than DLDR for all noise conditions and hence should provide better robustness to noise. In order to test whether higher GDLR levels predict better ASR accuracy, we evaluated correlation coefficient between the GDLR levels for

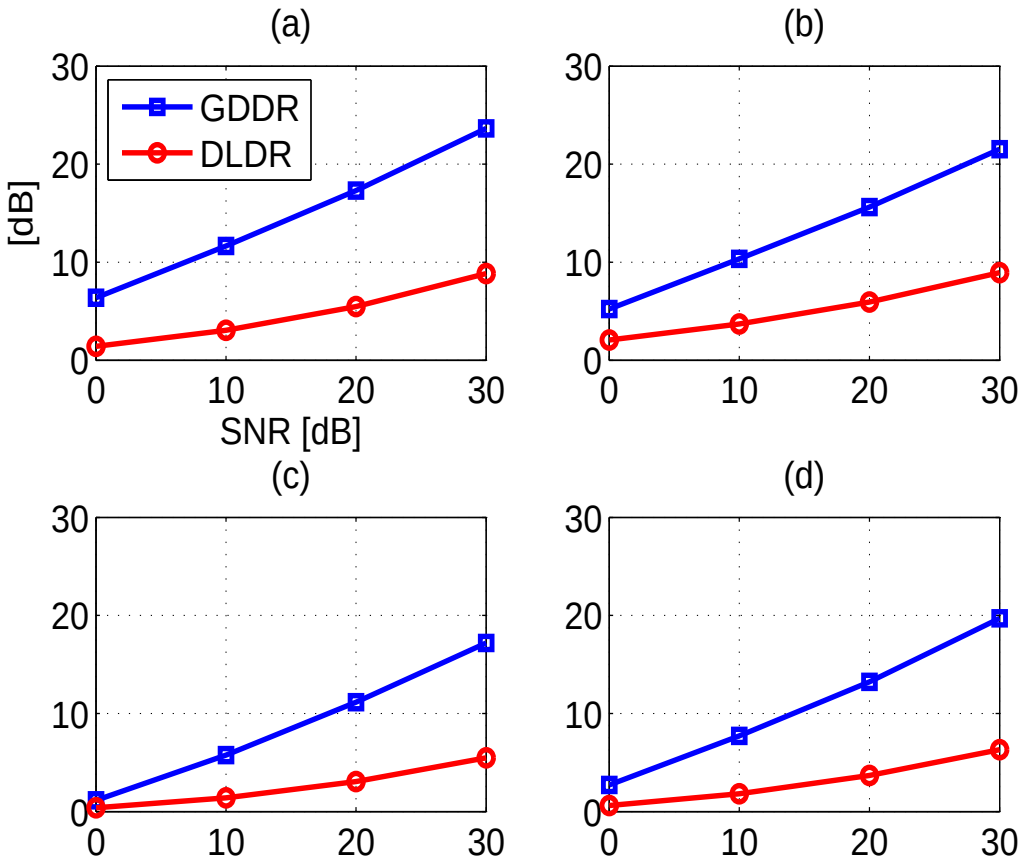


Fig. 7.8: *The GDDR and DLDR levels against SNR for additive noise conditions at* 10^{th} *Mel-channel in (a) White noise, (b) Background music, (c) Real-world noise, (d) Interfering speaker. Averaging was done over the Mel-channels.*

different noise types at 10 dB and the corresponding ASR accuracy which we found to be 0.78. Thus, the GDLR levels exhibit a good correlation with ASR accuracy.

7.5 EXPERIMENTAL RESULTS

We describe in this section experimental results comparing DSCC features to conventional MFCC/DCC and other features using degraded speech from the DARPA Resource Management (RM) database, which consists of 1600 training utterances and 600 test utterances. Data were obtained by digitally adding the various noises described above to the speech signal. We also evaluated the features in reverberant environments, which were simulated by convolving speech from the RM database with simulated room impulse responses using the (RIR) software package[1] [62]. Please see Sec. 4.3.1 for our ASR experimental setup.

7.5.1 The effect of d parameter in DSCC

Note that the DSCC features were obtained by applying the filter $H_d(z)$ in (7.2) to the speech Mel-Spectral sequences. In Fig. 7.10, we study the effect of the d parameter on DSCC features with respect to an ASR WER experiment. Fig. 7.9 plots the understandable frequency responses for the filter in (7.2) for the different d parameters. Observing Fig. 7.10, we note that $d = 3$ provides a better robustness to both noise and reverberation. Specifically for real-world noise conditions, DSCC with $d = 3$ provides a small but an additional 0.5-dB threshold shift than the DSCC with $d = 2$. The difference is more evident in presence of reverberation. Since reverberation leads to spectral smearing, hence it predominantly acts like a low-pass filter, a filter that has a better high-pass characteristics is more likely to compensate for reverberation, hence we understand that $d = 3$ should work better for reverberation. The parameter $d = 2$ provides better clean performance. Thus there is a trade-off between the clean performance and ASR robustness.

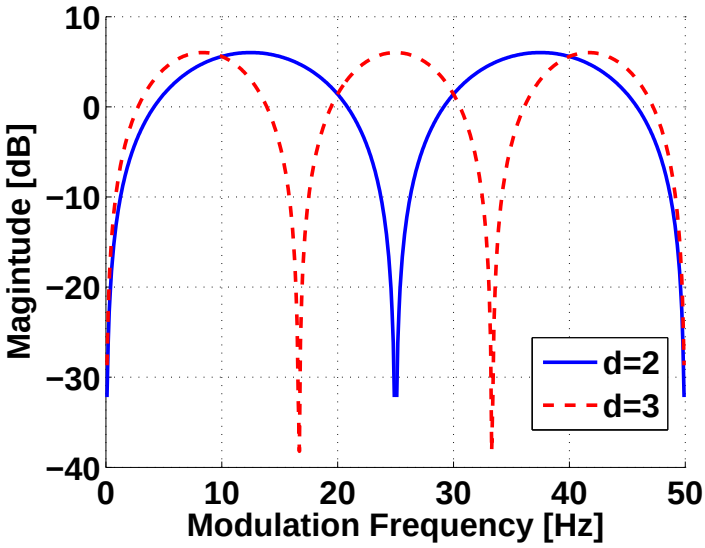


Fig. 7.9: *Frequency responses of the DSCC filter with different d parameters.*

7.5.2 DCC vs. DSCC

Fig. 7.11, compares the WER obtained using DCC, as in Fig. 7.1(a), against DSCC, where temporal-differencing is performed in the spectral domain,[2] as in Fig. 7.1(b). These comparisons clearly demonstrate the benefit of performing the time differencing in the spectral domain instead of in the conventional cepstral domain. It can be seen that the delta-spectral features substantially increase robustness to noise as well as reverberation, increasing the effective SNR by 5-8 dB at 50% WER. The use of DSCC features also provides a 30-45% relative reduction in WER at reverberation times of $300 - 500$ ms.

7.5.3 DSCC in Advanced Baseline Systems

Fig. 7.12 considers the combination of DSCC versus DCC features with MFCC, AFE [31] and PNCC [81], it can be seen that the use of the DSCC features provides better recognition accuracy than what is obtained from DCC features for all noise and

[1] `http://2pi.us/rir.html`

[2] The DSCC software is available at `http://www.cs.cmu.edu/~robust/archive/algorithms/DSCC_ICASSP2010/`.

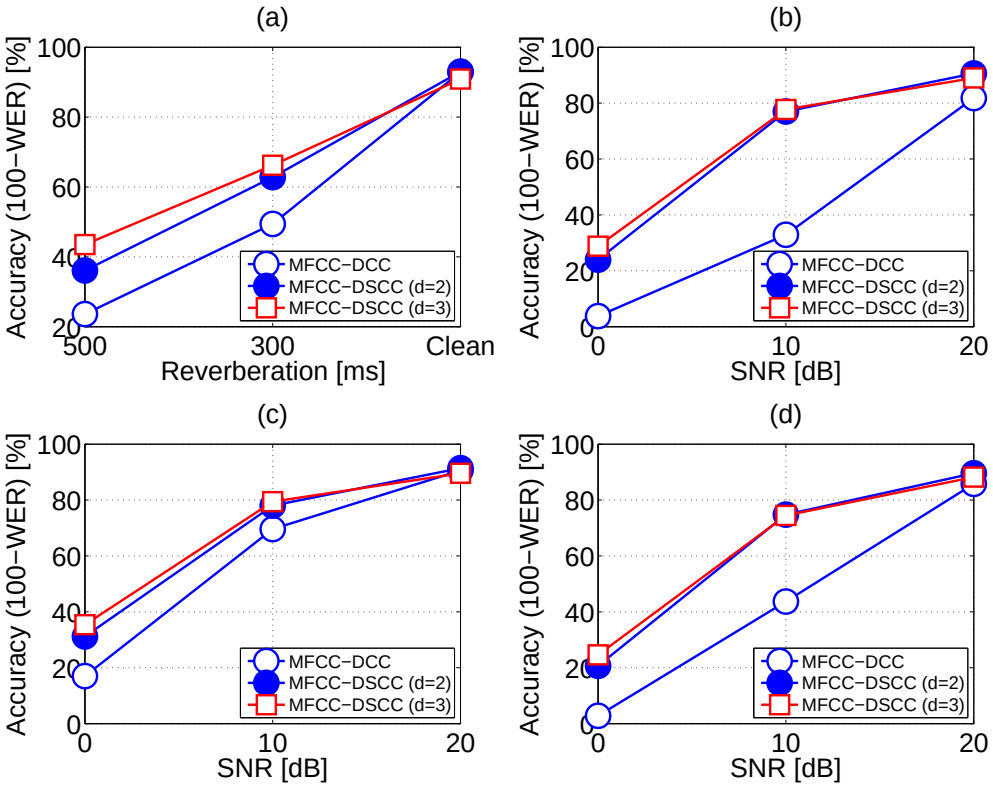
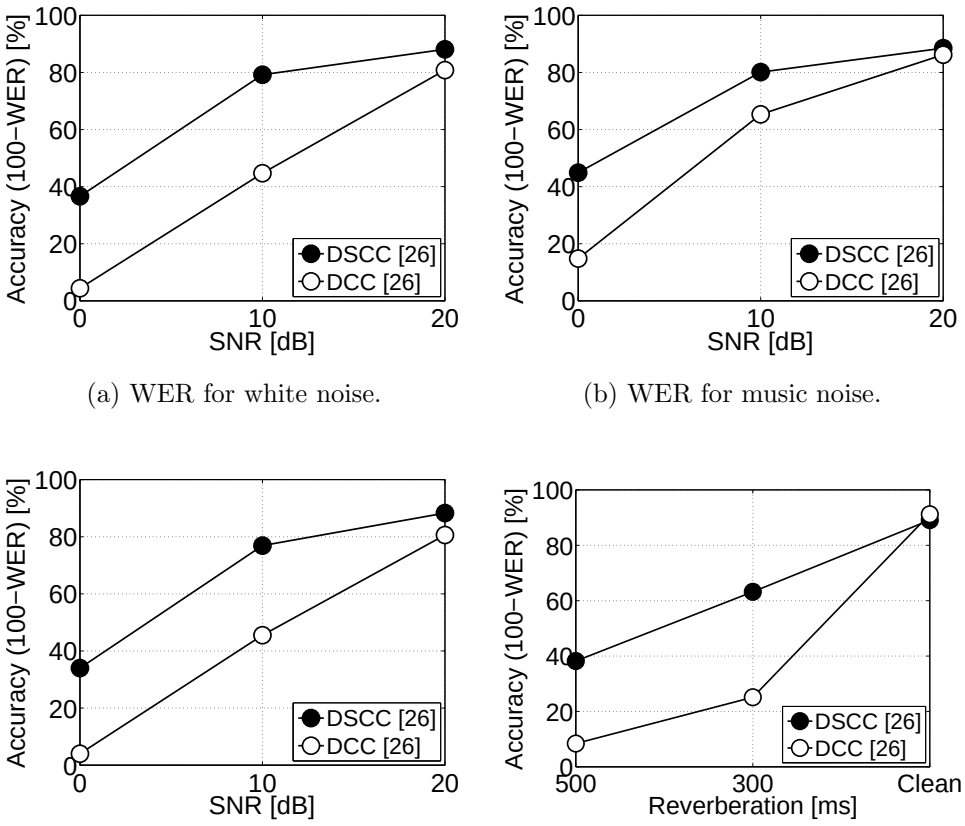


Fig. 7.10: *Word error rates (WERs) obtained in additive white noise using the d parameter in DSCC for (a) Reverberation, (b) White-noise, (c) Background music, (d) Real-world noise.*

reverberation conditions. The DSCC features not only strongly improve the baseline MFCC-DCC, they also improve the advanced systems in PNCC and AFE. Surprisingly we find that simply appending the 26-dim. DSCC features to the 13-dim. MFCC works as well as the conventional 39-dim. AFE features.

7.5.4 Magnitude Domain DSCC

In Fig. 7.13, we introduce a DSCC processing in the magnitude domain. This experiment was motivated by the results of magnitude domain vs. power domain NMF results in Ch. 6. There we found that magnitude domain NMF processing led to better ASR



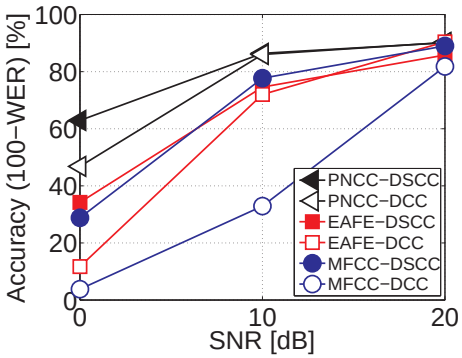
(a) WER for white noise.

(b) WER for music noise.

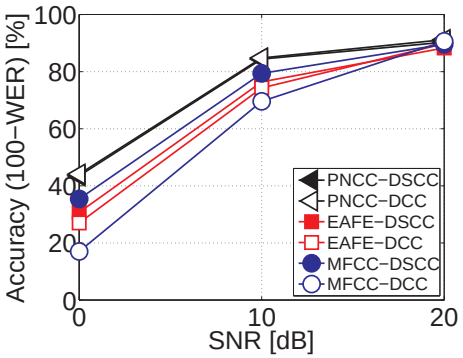
(c) WER for real-world noise recordings.

(d) WER for reverberation environments.

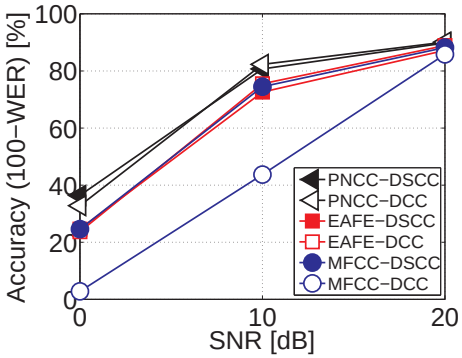
Fig. 7.11: *Comparisons of WERs for 26-dim. DCC and 26-dim. DSCC features in noisy and reverberant environments. MVN is included.*



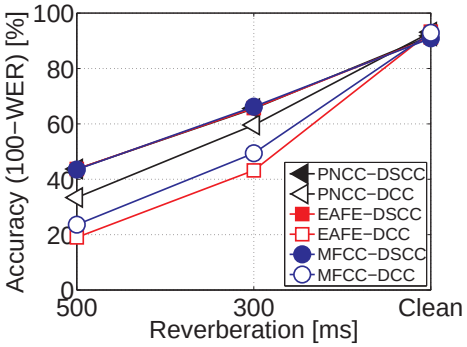
(a) WER for white noise.



(b) WER for music noise.



(c) WER for real-world noise recordings.



(d) WER for reverberation environments.

Fig. 7.12: *Comparisons of WERs obtained using DSCC versus DCC processing in combination with MFCC, PNCC, and AFE features. All the features are 39-dimensional and include MVN.*

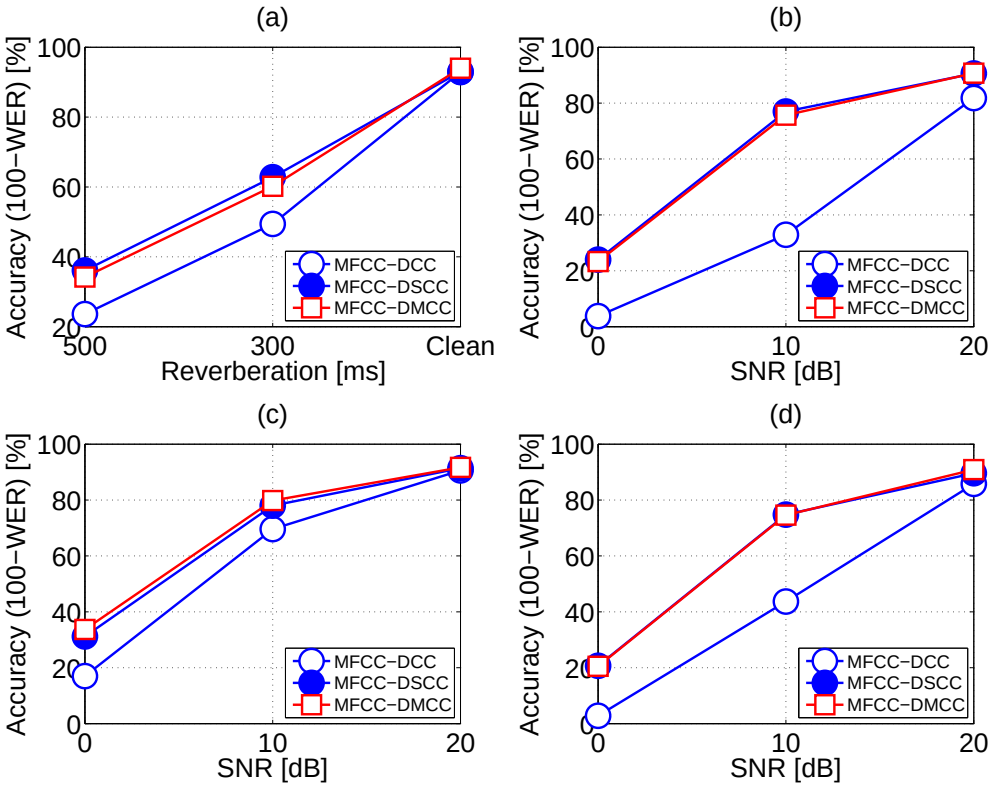


Fig. 7.13: *WERs obtained in magnitude domain DSCC for (a) Reverberation, (b) White Noise, (c) Background Music, (d) Real-World Noise.*

accuracy. In Fig. 7.13, we compare the DSCC results against DMCC, where DMCC was obtained from DSCC framework by replacing the power spectra with magnitude spectra. We see that DMCC and DSCC provide very similar robustness to ASR. A key difference between DMCC and DSCC lies in their clean performance. DSCC shows a relative loss of 4.5% in WER over the MFCC clean condition but DMCC shows an improvement of 10% over the corresponding MFCC features.

7.6 DISCUSSION

In this Chapter we proposed DSCC features that perform temporal differencing in the spectral domain rather than cepstral domain. We also find a good correspondence as a

function of noise type between the extent to which the use of DSCC processing reduces the WER and noise and the fraction of total noise power at DC. We also provided empirical results on spectral, log-spectral and cepstral distortion due to additive noise in the speech-signal domain. We also demonstrated that compared to DCC features, the DSCC features exhibit lower distortion levels and hence higher robustness to additive noise conditions. Overall, we observed that in comparison to conventional cepstral differencing, the use of DSCC features improves the effective SNR by 4-8 dB for various types of additive noise and reduces the relative WER by 20-30% in reverberation.

CHAPTER 8

A JOINT MODEL FOR NOISE AND REVERBERATION

In Chapter 3 we proposed a framework for representing reverberation in the spectral domain. In this chapter we provide a generalization of that model to include an additive noise term. We present our new model in Fig. 8.1, it serves the following key purposes:

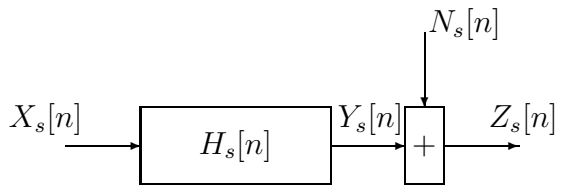


Fig. 8.1: *Modeling reverberation in spectral feature domain.*

1. The spectral domain model in Fig. 6.1 included an approximation error term that was then ignored. The model in Fig. 8.1 accounts for the approximation error in the sequence $N_s[n]$, and thus serves to extend the earlier model.
2. The model in Fig. 6.1 was a representation for only reverberation in the spectral domain. The new model generalized the earlier model by jointly modeling reverberation as well as noise in the spectral domain. The additive noise can be encapsulated by the $N_s[n]$ term in the new model.

Next we provide ASR experimental results using the new reverberation model in Fig. 8.1.

8.0.1 Joint NMF and DSCC processing for Dereverberation

In Fig. 8.2 we present dereverberation results using the joint NMF and DSCC processing, see Sec. 6.3 for experimental details. NMF is first applied on a reverberated

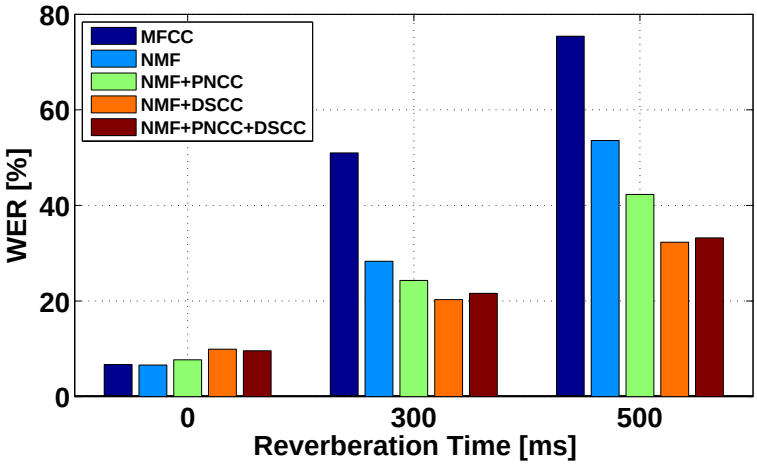


Fig. 8.2: *WERs for a joint NMF and DSCC processing.*

speech signal to dereverberate and then reconstruct the signal. DSCC then works on the NMF-reconstructed signal, where it replaces the traditional DCC features (see Ch. Ch:DSCC for details). Note that the inclusion of DSCC changes only the dynamic features, the static features remain unchanged. Under the joint framework, NMF is expected to compensate for the filter filter $H_s$in Fig. 8.1 and DSCC can compensate for the modeling error encapsulated by the additive term N_s. We see that at RT of 500 ms, the joint NMF and DSCC processing provides an additional 40% relative reduction over NMF processing and an overall 57% relative reduction in WER over MFCC. Similarly at RT of 300 ms the joint NMF and DSCC framework provides 60% relative reduction in WER over MFCC. Further, replacing MFCC by PNCC does not show additional improvements. We also experimented with adding LIFE processing in the NMF and DSCC framework but it did not provide substantial additional improvements.

8.0.2 Joint NMF and LIFE processing for Dereverberation

In Fig. 8.3 we study a joint NMF and LIFE processing framework on the DARRA RM-database. The degraded speech signal is first compensated by NMF, which results in a partially dereverberated speech signal. LIFE processing is subsequently applied to the NMF output. The combination of NMF and LIFE provides an overall 55% relative

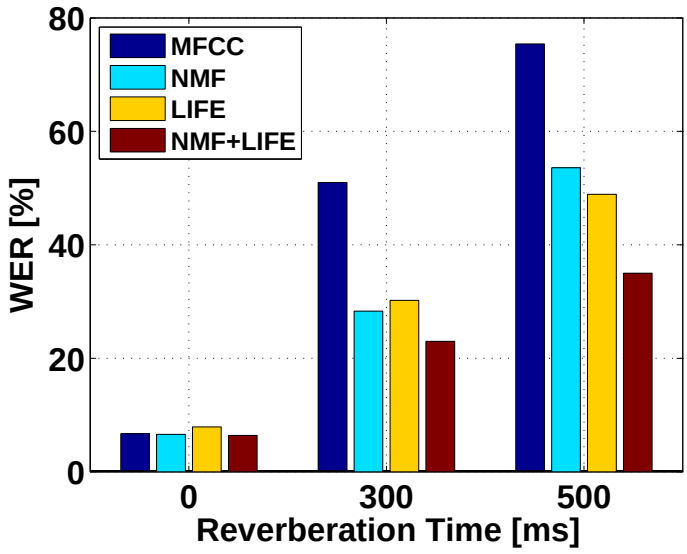


Fig. 8.3: *WERs for a joint NMF and LIFE processing.*

reduction in WER at RT of 300 ms, which is significantly better than 44% relative reduction with just NMF. Similarly at RT of 500 ms, the joint processing provides a 54% relative reduction in WER, whereas NMF alone provided only 30% relative reduction.

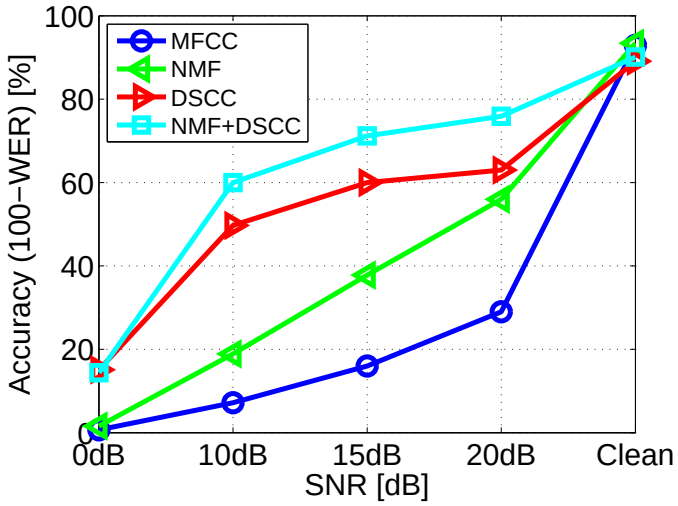


Fig. 8.4: *WERs for a joint noise and reverberation problem (RT of 300 ms).*

8.0.3 Joint Noise and Reverberation Compensation

In Fig. 8.4 we present a task that includes noise as well as reverberation, hitherto a very challenging task for ASR robustness. For this problem, the RIR was obtained at RT of 300 ms and noise was obtained from real-world noise recordings. We see that the baseline MFCC system nearly fails in this joint problem, with accuracy below 25% in any of the degraded environments. NMF, DSCC and their joint application according to the new model in Fig. 8.1 substantially improves the ASR performance. At 50% WER, DSCC processing provides a 12-dB horizontal threshold shift in equivalent SNR, whereas the joint NMF and DSCC processing provides 14-dB threshold shift.

8.1 DISCUSSION

In this chapter we proposed a framework for jointly representing reverberation and noise in the spectral domain. This framework generalizes the spectral domain reverberation framework in Chapter 3. We also provide experimental results on the potential combinations of NMF, DSCC, and LIFE processing. Among other tasks, we successfully demonstrated our approaches on a very challenging task that included both reverberation and noise.

CHAPTER 9

SUMMARY AND CONCLUSIONS

In this Chapter we provide a summary and key contributions of the thesis work. As stated before, the objective of this work was the development of signal processing and analysis techniques to sharply improve speech recognition accuracy in highly reverberant environments. We achieved our goal with a two-pronged approach. At first we studied and modeled the effects of reverberation directly in the speech-feature domain. Secondly, we provided compensation algorithms in the proposed framework. Next we provide a brief summary of the key contributions of this work.

9.1 MODELING REVERBERATION

In Chapter 3 we derived models for reverberation in the spectral, log-spectral and cepstral domains. During the spectral domain model derivation, we made an assumption in (3.10) that the energy in the cross-terms component is relatively small. Later in Chapter 8 we proposed a new reverberation model that encapsulated the approximation error as an additive noise. The new model serves as a generalization of the reverberation-only model as it provides a model for a joint noise and reverberation representation.

9.2 CEPSTRAL-POST FILTERING (CPF)

We proposed our first reverberation-compensation algorithm in terms of a cepstral-post filtering scheme in Chapter 4. The approach built on the framework in Chapter 3 to formulate and solve a least square problem for dereverberation in the cepstral domain. The approach required minimal assumptions on the reverberation filter coefficients and utilized speech knowledge in terms of it's cepstral auto-correlation sequences. Overall, CPF provided significant improvements in an ASR dereverberation task.

9.3 LIKELIHOOD-BASED INVERSE FILTERING (LIFE)

In Chapter 5 we provided a maximum-likelihood (ML) based approach for blind dereverberation. We theoretically motivated the ML approach in a simplified setting. LIFE achieved dereverberation in the cepstral domain by formulating and solving a ML criterion based on the probability distribution of speech cepstral features. LIFE processing provided 40-45% relative reduction in WER under clean-training conditions. The LIFE processing benefits extended to multi-style training as well, where it provided 12-22% relative reduction in WER. LIFE processing also showed 18% relative reduction in WER on top of supervised MLLR.

9.4 NON-NEGATIVE MATRIX FACTORIZATION (NMF)

In chapter 6 we proposed NMF for speech dereverberation in the spectral domain. NMF is based on a mean-squared error optimization which builds on the non-negativity and sparsity of the spectral values. NMF is also a blind approach for dereverberation as the algorithm must work blindly for the different room reverberation conditions. We showed that applying NMF on magnitude spectra, rather than on power spectra, provided superior performance. We also integrated perceptual weighting in the NMF framework to enhance the speech frequencies from an auditory perspective. We experimentally showed that NMF provides up to 40-45% relative reduction in WER for clean-training. The NMF benefits extend also to multi-style training.

9.5 DELTA-SPECTRAL CEPSTRAL COEFFICIENTS (DSCC)

In Chapter 7 we provided a signal processing scheme for noise robustness, where we proposed DSCC features that perform temporal differencing in the spectral domain rather than cepstral domain. We also proposed a suitable non-linearity on the delta-spectra features for their application in speech recognition. We performed distortion analysis of the delta-spectra, Gaussianized-delta-spectra and derived cepstral features. We showed that the approach provided significant improvement for stationary as well as non-stationary noises.

9.6 COMPARATIVE OVERVIEW OF ALGORITHMS

In Table 9.1 we briefly provide a comparative overview of the algorithms presented in the thesis. We compare the algorithms with respect of their computational cost, oracle performance, requirement of *a prior* room reverberation knowledge and the feasibility of their online implementation. We represent computational cost for different algorithms in terms of that required in MFCC feature extraction, which we represent as 1x. ASR involves feature decoding on top of feature extraction, so for better perspective on computational requirements, we provide computational cost for ASR decoding. We see that compared to MFCC, CPF and DSCC are relatively inexpensive algorithms. LIFE and NMF algorithms are more expensive than MFCC but if we incorporate ASR decoding computations, LIFE is only 20% more expensive than MFCC, whereas NMF is about 90% more expensive than MFCC.

	Computation	Oracle experiment	*A priori* room knowledge	Online
MFCC	1.0x	-	-	Yes
CPF	0.1x	-	None	Yes
LIFE	10x	11% WER at RT of 500 ms	None	Yes
NMF	45x	21% WER at RT of 500 ms	None	No
DSCC	1.2x	-	None	Yes
ASR decoding	50x	-	-	-

Table 9.1: *A comparative overview of algorithms.*

We compared LIFE and NMF under an oracle experiment in Sec. 6.3.1 and found that LIFE processing showed stronger potential than NMF. Our algorithms do not require any *a prior* knowledge about RIR, source/speaker location etc. and are completely blind to the room conditions. CPF, LIFE and DSCC algorithms can be made online for stationary or slowly-changing environments. The required reverberation compensation parameters can be estimated from past utterances, which can be used for current utterance and re-estimated for future utterances. The reverberation com-

pensation stage for these algorithms can then be done online. NMF on the other stage is a special algorithm, it's dereverberation stage works in an optimization framework that requires statistics from the RIR as well as the utterance under consideration. Correspondingly the NMF reverberation compensation can not be made online.

REFERENCES

[1] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[2] J. F. Culling, Q. Summerfield, and D. H. Marshall, "Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels," *Speech Communication*, vol. 14, pp. 71–95, 1994.

[3] C. J. Darwin and R. W. Hukin, "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," *Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 335–342, 2000.

[4] J. F. Culling, K. I. Hodder, and C. Y. Toh, "Effects of reverberation on perceptual segregation of competing voices," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2871–2876, 2003.

[5] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[6] S. G. McGovern, "A model for room acoustics," http://2pi.us/rir.html.

[7] D. R. Campbell, K. J. Palomäki, and G. Brown, "A MATLAB simulation of shoebox room acoustics for use in research and teaching," *Computing and Information Systems Journal, ISSN 1352-9404*, vol. 9, no. 3, 2005.

[8] S. A. Gelfand and S. Silman, "Effects of small room reverberation upon the recognition of some consonant features," *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 22–29, 1979.

[9] P. J. Bloom, "Evaluation of a dereverberation process by normal and impaired listeners," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Denver, April 1980, pp. 500–503.

[10] S. Devore, B. G. Shinn-Cunningham, N. I. Durlach, and H. S. Colburn, "The influence of reverberation on spatial release of masking in consonant identification," *Journal of the Acoustical Society of America*, vol. 111, p. 2422, 2002.

[11] K. Eneman, J. Duchateau, M. Moonen, D. V. Campernolle, and H. V. Hamme, "Assessment of dereverberation algorithms for large vocabulary speech recognition systems," in *Proc. of Eurospeech*, Geneva, September 2003, pp. 2689–2692.

[12] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr. 1988, pp. 651 –654.

[13] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene database in real acoustic environments," in *Oriental COCOSDA Workshop*, 1998.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[15] R. Singh, R. M. Stern, and B. Raj, "Signal and feature compensation methods for robust speech recognition," *Noise Reduction in Speech Applications, Ed. G. Davis, Chapter 9*, pp. 221–246, 2002, CRC Press LLC, USA.

[16] R. Singh, B. Raj, and R. M. Stern, "Model compensation and matched condition methods for robust speech recognition," *Noise Reduction in Speech Applications, Ed. G. Davis, Chapter 10*, pp. 247–278, 2002, CRC Press LLC, USA.

[17] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustic Society*

of America, pp. 1304–1312, 1974.

[18] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. of the International Conference on Spoken Language Processing*, 2002, pp. 2185–2188.

[19] A. Sehr and W. Kellermann, "A new concept for feature-domain dereverberation for robust distant-talking asr," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. IV–369–IV–372.

[20] ——, "Model-based dereverberation of speech in the mel-spectral domain," in *Proc. IEEE Asilomar*, 2008, pp. 783–787.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[22] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on SAP*, vol. 8, pp. 267–281, May 2000.

[23] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 3701–3704.

[24] S. Ganapathy, T. Samuel, and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech," *Journal of the Acoustical Society of America, Express Letters*, Jan. 2009.

[25] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Acoustics*, vol. 2, pp. 587–589, Oct. 1994.

[26] R. H. Kay, "Hearing of modulation in sounds," *Physiological Reviews*, vol. 62, pp. 894–975, 1982.

[27] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1–3, pp. 117–132, 1998.

[28] B. E. D. Kingsbury, "Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments," Ph.D. dissertation, University of California, Berkeley, 1998.

[29] Y.-H. B. Chiu and R. M. Stern, "Minimum variance modulation filters for robust speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[30] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. Interspeech*, 2009.

[31] ETSI: Advanced Front-end, ETSI Doc. No. ES 202 050.

[32] D. Dimitriadis, J. C. Segura, L. Garcia, R. Potamianos, P. Maragos, and V. Pitsikalis, "Advanced front-end for robust speech recognition in extremely adverse environments," in *Proc. Interspeech*, 2007.

[33] N. Cahill and R. Lawlor, "A novel approach to mixed phase room impulse response inversion for speech dereverberation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4593–4596.

[34] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on ASLP*, vol. 17, no. 4, pp. 534–545, May 2009.

[35] M. Wu and D.-L. Wang, "A Two-Stage Algorithm for Enhancement of Reverberant Speech," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 1085–108.

[36] E. Habets, S. Gannot, I. Cohen, and P. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1433–1451, Nov. 2008.

[37] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising us-

ing multichannel linear prediction," , *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

[38] H. Gölzer and M. Kleinschmidt, "Importance of early and late reflections for automatic speech recognition in reverberant environments," in *Elektronische Sprachsignalverarbeitung (ESSV)*, 2003.

[39] A. K. Nabelek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *Journal of the Acoustical Society of America*, vol. 71, pp. 1242–1248, 1982.

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

[41] 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 *Proc. Interspeech*, 2009.

[42] R. M. Stern, E. B. Gouvêa, and G. Thattai, "Polyaural array processing for automatic speech recognition in degraded environments," in *Proc. of Interspeech*, September 2007.

[43] E. Habets, "Single and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, TU Eindhoven, 2007.

[44] M. L. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Dept. of ECE, Carnegie Mellon University, 2003.

[45] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant environments," *Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 4040–4047, December 2006.

[46] J. Blauert, *Spatial Hearing*, 2nd ed. Cambridge, MA: MIT Press, 1996.

[47] G. J. Brown and K. J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis*, G. Brown and D. Wang, Eds. Wiley and IEEE Press, 2006.

[48] H.-M. Park and R. M. Stern, "Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings," *Speech Communication*, pp. 15–25, 2009.

[49] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Tran. on ASLP*, vol. 15, No.7, pp. 2023–2032, 2007.

[50] Q. Jin, K. Kumar, R. M. Stern, and T. Schultz, "Speaker identification of simulated far-field speech," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2008, submitted.

[51] A. Acero, and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Apr. 1990, pp. 849–852.

[52] A. Acero, "Acoustical and environmental robustness for automatic speech recognition," Ph.D. dissertation, Dept. of ECE, Carnegie Mellon University, 1990.

[53] F. H. Liu, R. M. Stern, and A. A. and P. J. Moreno, "Environment normalization for robust speech recognition using direct cepstral comparison," in *Proc. International Conference Acoustics, Speech, Signal Processing*, vol. 2, Apr. 1994, pp. 61–64.

[54] K. Kumar and R. M. Sterm, "Environment-invariant compensation for reverberation using linear post-filtering for minimum distortion," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[55] J. Campbell and D. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[56] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. of the IEEE*, vol. 85, pp.

1437–1462, september 1997.

[57] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, August.

[58] J. S. Gammal and R. A. Goubran, "Combating reverberation in speaker verification," *Instrumentation and Measurement Technology Conference*, 2005.

[59] B. Xiang, U. V. Chaudhari, J. Navratil, G. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 681–684.

[60] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1993.

[61] F. Jelinek, *Statistical Methods for Speech Recognition (Language, Speech, and Communication).* MIT Press, 1998.

[62] K. Kumar and R. M. Sterm, "Maximum-likelihood-based cepstral inverse filtering," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[63] J. M. Mendel, *Maximum-likelihood deconvolution: A journey into model-based signal processing.* Sprinter-Verlag, 1990.

[64] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 45–48.

[65] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," in *Proc. InterSpeech*, 2009, pp. 1231–1234.

[66] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. International Conference on Acous-*

tics, Speech and Signal Processing, Orlando, May 2002, pp. 557–560.

[67] M. J. Gales and P. C. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech and Language*, vol. 10, pp. 249–264, Oct. 1996.

[68] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book 3.2," Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/.

[69] K. Kumar, R. Singh, B. Raj, and R. M. Sterm, "Gammatone sub-band magnitude-domain dereverberation for asr," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[70] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, 1997.

[71] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. ICA*, 2004, pp. 494–499.

[72] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling." in *Proc. Neural Information Processing Systems Workshop on Advances in Models for Acoustic Processing*, 2006.

[73] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, "History and future of auditory filter models," in *Proc. ISCAS*, 2010.

[74] K. Kumar, R. Singh, B. Raj, and R. M. Sterm, "An iterative least-squares technique for dereverberation," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[75] K. Kumar, "A spectro-temporal framework for compensation of reverberation for speech recognition," Ph.D. Proposal, Dept. of ECE, Carnegie Mellon University, 2010, http://www.ece.cmu.edu/~kshitizk/Thesis/Proposal.pdf.

[76] X. Huang, A. Acero, and H.-W. Won, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Upper Saddle River, NJ, USA:

Prentice-Hall, Inc., 2001.

[77] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on ASSP*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[78] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997.

[79] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1996.

[80] H. G. Hirsch, P. Meyer, and H. W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. of Eurospeech*, Genoa, September 2001, pp. 413–416.

[81] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[82] S. Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 1986.

[83] P. M. Zurek, *The precedence effect, Ch. 4*. New York, NY: Springer-Verlag, 1987.

[84] D. M. Green, *An Introduction to Hearing*, 6th ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1976.

[85] J. Chen, K. K. Paliwal, and S. Nakamura, "Cepstrum derived from differentiated power spectrum for robust speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 469–484, Oct 2003.