

Multisensor Segmentation-based Noise Suppression for Intelligibility Improvement in MELP Coders

A Thesis
Presented to
The Academic Faculty

by

Cenk Demiroglu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2006

Multisensor Segmentation-based Noise Suppression for Intelligibility Improvement in MELP Coders

Approved by:

Professor Mark A. Clements, Committee
Chair

Professor Paul Hasler

Professor David V. Anderson, Adviser

Professor Levent Degertekin
(Mechanical Engineering)

Professor Thomas Barnwell III

Date Approved: December 12, 2005_____

Dedicated to Rahman, the most Merciful...

ACKNOWLEDGEMENTS

I would like to thank God for making this dissertation exist. The graduate study has been a long journey with many academic and mostly nonacademic challenges. Each time I found myself in trouble, I felt his hand reaching out and holding me really tight. Without his help and support, this dissertation could only be a dream for me.

One of the turning points in my graduate studies, as well as my life, is when I started working with my advisor Dr David V. Anderson. I knew he was a great guy from the very beginning, but I was not expecting the wonderful years I spent in his lab. I believe that it was God's blessing to meet somebody like David. I still remember the days I was working with him in his office. One day he looked at the sky and said "Sky is the limit." I thought he was kidding me. He was not... His strong support, friendliness, scientific and spiritual inspiration made me throw away the phd blues and start a breakthrough period in my life. He gave me the opportunity to discover and improve my intellectual abilities with a patience I probably would not have if I was in his place. He is undoubtedly one of the most influential people in my life in many ways, and I am thankful to him for everything he did for me.

If there is one person besides David that I cannot forget in this acknowledgement section, it is Dr Mark A. Clements. I even paid attention to putting the "A." in his name because of dedication to details that I became aware of, cannot really say learned, thanks to Mark. His pattern recognition class was the most inspiring class I ever took. I asked him hundreds of questions, and I do not remember a single time that he was tired of sharing his knowledge with me in the most inspiring and energetic ways. Early in my phd study, I remember one of the CSIP parties where I saw his energy, his caring, and positive attitude towards all the students he was talking to. I thought life would be much better if I am around him more often. He did not prove me wrong. Thanks to God's mercy, I got the opportunity to work with him and learned many things from this rare type of professor who is always ready to

educate his students both by talking to them and by setting an example to them. I admit that I still could not solve many things about Mark, but I am thankful to everything I learned from him in the last four years.

One of the strange things that happened to me during my phd study was working with Dr Thomas Barnwell. I could only see him in the weekly meetings but the strange thing is that he has such an influence in this dissertation. One of the best things I learned from him is the divide-and-conquer approach he took to solve the speech problems. I noticed his way in the meetings, and I realized that dividing the problem into subproblems instead of taking a holistic approach made solving the speech problems easier. Although I still think that the ultimate solution should be in a holistic framework, the bottom-up approach certainly opened the way for progress and breakthroughs in my mind. He is also the one who made me look at the speech enhancement problem from the auditory filterbank perspective. I was talking to him one day and I noticed how natural it became for him to think about speech perception from a filterbank point of view. Strangely enough, two minutes of talk that day created a growing curiosity in my mind that led to acquiring a vast amount of knowledge in the psychoacoustics field. As an outcome of his influence, I now have a personal library at home with a focus on acoustic-phonetics and psychoacoustics. I even read the great book "The Dream Machine" thanks to him mentioning it to us in a group meeting. The influence he has on me and some of the fundamental things I learned from him are invaluable.

My labmates were really patient with me although I kept pulling their legs, asking them questions in the meetings, and bugging them all the time. I want to thank all the unnamed heroes in my lab for being such beautiful people with such wonderful hearts. Without their friendship, this dissertation would still be possible, but they made it so much easier for me to go through the last four years. They became like brothers and sisters who will unfortunately have to be in touch with me for the rest of their lives.

I would like to thank my mother and father partly because it is the tradition to do so and mainly because God will be happy with me. I know that they had a really hard time back at home when their son is so far away, though I tend to see it as a gift from God. My mom discovered internet and MSN, and made sure I am emotionally abused at all times.

Without her MSN messages and emails, life would be more boring but less depressing. Still, knowing that they love me, gave me the motivation to complete my studies and make them proud of me which they are. I am thankful to God for blessing me with such loving and caring parents.

Many friends, girlfriends, roommates, and other uncategorized folks became part of my life in the last four years. Some of them made me happy, some of them made me angry and even upset me. But they all served a purpose in my life and help me socialize, get to know life and people better. I believe they all deserve an acknowledgement, and I want to take this opportunity to thank all of them for being part of my life and make my phd life a relatively humane one.

TABLE OF CONTENTS

| | |
|--|-------------|
| DEDICATION | iii |
| ACKNOWLEDGEMENTS | iv |
| LIST OF TABLES | xi |
| LIST OF FIGURES | xiii |
| SUMMARY | xv |
| I INTRODUCTION | 1 |
| 1.1 Motivation of Research | 1 |
| 1.2 Scope of Thesis | 1 |
| 1.3 Organization of Thesis | 3 |
| II BACKGROUND | 4 |
| 2.1 Speech Enhancement | 4 |
| 2.1.1 Single Microphone Statistical Speech Enhancement | 5 |
| 2.1.2 Multi-Sensor Speech Enhancement | 10 |
| 2.1.3 Noise Estimation | 11 |
| 2.1.4 SNR Estimation | 11 |
| 2.1.5 Assessing the Performance of a Speech Enhancement System | 12 |
| 2.2 Noise-Robust Automatic Speech Recognition | 13 |
| 2.2.1 Missing Data Techniques | 14 |
| 2.2.2 Noise-Robust Feature Extraction | 14 |
| 2.2.3 Speech Enhancement for Noise-Robust ASR | 17 |
| 2.2.4 Feature Enhancement | 17 |
| 2.3 Noise-Robust Speech Segmentation and Classification | 18 |
| 2.3.1 Discriminative Feature Extraction | 18 |
| 2.3.2 HMM-based Techniques | 18 |
| 2.3.3 Spectral Variation-based Techniques for Segmentation | 19 |
| 2.4 Description of the GEMS Device and the ARCON Database | 19 |
| 2.4.1 ARCON Database Description | 20 |

| | | |
|------------|--|-----------|
| 2.4.2 | Correlation of GEMS-based and Acoustic Microphone-based Features | 20 |
| 2.4.3 | Time Delay Between the Acoustic Sensor and the GEMS Sensor . . . | 23 |
| III | A PERCEPTUALLY-MOTIVATED BAYESIAN RISK MINIMIZATION FRAMEWORK FOR SPEECH ENHANCEMENT | 25 |
| 3.1 | Proposed Framework | 26 |
| 3.1.1 | Sources of Misestimations in Speech Enhancement Filters | 28 |
| 3.1.2 | Costs of Misestimations | 28 |
| 3.1.3 | Quantifying the Probability of Misestimation | 29 |
| 3.2 | Analysis of Spectral Subtraction Method Using the Proposed Framework . | 32 |
| 3.2.1 | Analysis of the Modified Spectral Subtraction | 32 |
| 3.2.2 | Spectral Subtraction for Perceptual Quality Enhancement | 32 |
| 3.2.3 | Spectral Subtraction for Noise-Robust ASR | 35 |
| IV | AHARMONIC COMB FILTERING OF NOISY SPEECH | 37 |
| 4.1 | Aharmonic Comb Filter | 38 |
| 4.1.1 | The HSP Detection Problem | 38 |
| 4.1.2 | Detection of HSP Locations in Voiced Speech Spectrum Using the GEMS Device | 39 |
| 4.2 | Speech Enhancement Using the Proposed HSP Detector | 42 |
| 4.3 | Experiments | 43 |
| 4.3.1 | Results | 44 |
| V | SPARSE REPRESENTATION OF VOICED SPEECH SPECTRUM FOR NOISE-ROBUST LPC EXTRACTION | 46 |
| 5.1 | Effect of Sparse Spectrum on LPC Extraction | 47 |
| 5.1.1 | Noise-robust LPC Extraction Using the Aharmonic Comb Filter . . | 48 |
| 5.2 | Experiments | 49 |
| 5.3 | Conclusion | 51 |
| VI | SEGMENTATION-BASED ENHANCEMENT OF THE NOISY SPEECH SIGNAL FOR INTELLIGIBILITY IMPROVEMENT IN MELP SPEECH CODERS | 52 |
| 6.1 | Problem Definition | 54 |
| 6.2 | Auxiliary Sensors | 55 |
| 6.3 | Proposed System | 56 |

| | | |
|---|--|-----------|
| 6.3.1 | The Segmentation Algorithm | 57 |
| 6.3.2 | The MMSE Estimator | 58 |
| 6.4 | Methods Used in Signal Presence Detection | 59 |
| 6.5 | Sound Class Dependent Signal Presence Detection | 60 |
| 6.5.1 | Vocalics | 60 |
| 6.5.2 | Unvoiced Fricatives | 60 |
| 6.5.3 | Unvoiced Plosives | 61 |
| 6.5.4 | Voiced Plosives | 61 |
| 6.5.5 | Voiced Fricatives | 63 |
| 6.5.6 | Transitional Sounds and Transients | 63 |
| 6.5.7 | Silence | 63 |
| 6.6 | Experiments | 64 |
| 6.6.1 | Description of the Diagnostic Rhyme Test (DRT) | 65 |
| 6.6.2 | Analysis of Segmentation Errors | 65 |
| 6.6.3 | Results | 66 |
| 6.7 | Summary and Conclusions | 75 |
| VII USING A SIMULATED VOICING FEATURE FOR NOISE-ROBUST ASR | | 77 |
| 7.1 | Three Strategies for Utilizing the Voicing Feature | 79 |
| 7.1.1 | Feature Fusion Method | 79 |
| 7.1.2 | Analysis of the SNR-Dependent Performance Gain with the Voicing Feature Method | 79 |
| 7.1.3 | Data Marginalization Method | 81 |
| 7.1.4 | Proposed System | 82 |
| 7.1.5 | Voicing Extraction Algorithm | 82 |
| 7.1.6 | High-Energy Speech Detection | 83 |
| 7.2 | Simulating Voicing and High-Energy Speech Detection | 84 |
| 7.2.1 | Simulating the GEMS-based Voicing Detection | 84 |
| 7.2.2 | Simulating the GEMS-based High-Energy Speech Detection | 85 |
| 7.2.3 | Incorporating the Simulation Errors into ASR Experiments | 86 |
| 7.3 | Experiments | 87 |

| | | |
|-------------|---|------------|
| 7.3.1 | Experimental Setup | 87 |
| 7.3.2 | Results | 88 |
| VIII | EFFECTS OF SEVERE SENSOR NOISE ON GEMS-BASED FEAT- TURES | 94 |
| 8.1 | Noise Interference in the GEMS Signal | 94 |
| 8.2 | Feature-Fusion Based ASR with Noisy GEMS Signal | 97 |
| 8.3 | Experiments | 98 |
| 8.3.1 | Results and Discussion | 99 |
| IX | CONCLUSION | 103 |
| | REFERENCES | 106 |
| | VITA | 116 |

LIST OF TABLES

| | | |
|----|---|----|
| 1 | The parameters that are used in analysis of speech and GEMS signals in this chapter. | 23 |
| 2 | Results of the A/B speech quality tests. | 44 |
| 3 | 95 percent confidence intervals of the A/B test results using student's t-test. | 44 |
| 4 | Comparison of the proposed system with the noisy system using the segmental SNR measure for the M2 tank noise. | 44 |
| 5 | Comparison of the proposed system with the noisy system using the log-likelihood ratio measure for the M2 tank noise. | 45 |
| 6 | Comparison of the proposed system with the noisy system using the segmental SNR measure for the Blackhawk helicopter noise. | 45 |
| 7 | Comparison of the proposed system with the noisy system using the log-likelihood ratio measure for the Blackhawk helicopter noise. | 45 |
| 8 | Comparison of the Itakura-Saito distortion measure for noisy speech with full-spectrum and with sparse spectrum. | 51 |
| 9 | Comparison of the Itakura-Saito distortion measure for noisy speech with full-spectrum and with sparse spectrum. | 51 |
| 10 | Combinations of the acoustic-phonetic, HSP location, and <i>SNR</i> information that are used for signal presence detection are illustrated for each sound class. | 58 |
| 11 | Parameter values used by the proposed speech enhancement system. | 64 |
| 12 | The DRT scores for each distinctive feature compared for MELP encoded speech in 0 <i>dB</i> M2 noise environment. | 66 |
| 13 | Detection rates (%) of the four consonant sounds using the segmentation algorithm. | 67 |
| 14 | Mutual Confusion rates (%) among four consonant classes using the segmentation algorithm. | 67 |
| 15 | Averaged word error rate (WER) results are shown using the data marginalization method with various energy threshold levels. | 85 |
| 16 | Probability of error is shown for high-energy speech detection using the GEMS device for various energy thresholds. | 87 |
| 17 | Performance results in terms of WER for four noise types using the baseline system. | 90 |
| 18 | Performance results in terms of WER for four noise types using the DM method. | 91 |

| | | |
|----|---|-----|
| 19 | Performance results in terms of WER for four noise types using the data fusion method. | 91 |
| 20 | Performance results in terms of WER for four noise types using the proposed system. | 92 |
| 21 | Performance results in terms of WER for four noise types using the standard missing data system. | 92 |
| 22 | Performance results in terms of WER for four noise types using the 24 Mel-spaced filter outputs. | 92 |
| 23 | ASR performance when a GEMS-based feature (energy or periodicity) is fused with the MFCC feature vector. | 102 |
| 24 | GEMS signal is enhanced with an MMSE estimator. ASR performance when a GEMS-based feature (energy or periodicity) is fused with the MFCC features. | 102 |
| 25 | A missing data technique is applied to the GEMS-based feature. ASR performance when a GEMS-based feature (energy or periodicity) is fused with the MFCC features. | 102 |
| 26 | The quality and intelligibility tests that have been done to measure the performance of the proposed enhancement systems. | 104 |

LIST OF FIGURES

| | | |
|----|---|----|
| 1 | Spectrograms of a sample speech signal and simulatenously recorded GEMS signal are shown with the trajectory of the periodicity features. | 21 |
| 2 | Trajectory of GEMS-based and speech-based energy features are compared. | 22 |
| 3 | Magnitude spectrums of the GEMS signal and acoustic signal are compared for a voiced speech frame. | 24 |
| 4 | Effects of oversuppression and undersuppression is illustrated for noisy speech using spectrogram analysis. | 30 |
| 5 | The gain functions H (dB) is shown in the top figure. Sensitivity of the gain functions to the prior parameters are shown in the bottom figure. | 34 |
| 6 | A sample voiced speech spectrum. | 39 |
| 7 | An illustration of the HSP detection algorithm. | 41 |
| 8 | System overview of the aharmonic comb filter. | 42 |
| 9 | Overview of the proposed LPC extraction system. | 49 |
| 10 | Illustration of noise-robust LPC extraction with the proposed system. . . . | 50 |
| 11 | Spectrograms of the GEMS and p-mic signals are compared. | 56 |
| 12 | An overview of the proposed segmentation-based speech enhancement system. | 57 |
| 13 | Power spectrum of a 20 msec speech sample taken from the bust location of a plosive sound is shown for clean and noisy speech. SNR of the noisy spectrum is also shown. | 62 |
| 14 | Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the voicing feature. | 68 |
| 15 | An example case is shown where the segmentation error causes confusion in voicing. | 69 |
| 16 | Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. | 70 |
| 17 | Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. | 71 |
| 18 | Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. | 72 |

| | | |
|----|---|-----|
| 19 | Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in nasality. . . . | 73 |
| 20 | Spectrograms of the M-MMSE algorithm and the proposed system are compared to show the improvement in sibilant for a monosyllable sample. . . | 74 |
| 21 | Energy trajectories of the clean speech and the noisy speech are compared, and energy thresholding is illustrated. | 82 |
| 22 | Flow diagram of the proposed system is shown. | 83 |
| 23 | Probability of error for high-energy speech detection using the GEMS device is shown. | 86 |
| 24 | Energy trajectories of the clean speech signal and the radar signal are shown. Energy threshold of -5 dB is used for both cases, and the high-energy segments are shown with rectangular boxes. | 87 |
| 25 | Probability of error is shown for the simulation of voicing detection. | 88 |
| 26 | Performance improvement over the baseline system using the feature fusion (FF) method, data marginalization (DM) method, and the proposed system. | 89 |
| 27 | Comparison of the performance between the proposed system, the baseline system, and other relatively low complexity noise-robust systems are shown. | 93 |
| 28 | Spectrogram of a clean speech signal is shown in the top figure. Spectrogram of the corresponding noisy GEMS signal is shown in the bottom Figure. . . | 95 |
| 29 | Sample spectrograms of a GEMS signal and the corresponding audio signal are shown with the trajectory of the voicing features. | 96 |
| 30 | Energy trajectory of a sample speech utterance and energy trajectory of the GEMS signal are shown. | 97 |
| 31 | Magnitude spectrums of a noisy GEMS signal and acoustic signal are compared for a voiced speech frame. | 98 |
| 32 | Spectrogram of a severely noisy GEMS sample. | 101 |
| 33 | Comparison of the pdf of energy features extracted from all (noisy and clean) GEMS utterances and severely noisy GEMS utterances are shown. | 101 |

SUMMARY

This thesis investigates the use of an auxiliary sensor, the GEMS device, for improving the quality of noisy speech and designing noise preprocessors to MELP speech coders. Use of auxiliary sensors for noise-robust ASR applications is also investigated to develop speech enhancement algorithms that use acoustic-phonetic properties of the speech signal.

A Bayesian risk minimization framework is developed that can incorporate the acoustic-phonetic properties of speech sounds and knowledge of human auditory perception into the speech enhancement framework. Two noise suppression systems are presented using the ideas developed in the mathematical framework. In the first system, an aharmonic comb filter is proposed for voiced speech where low-energy frequencies are severely suppressed while high-energy frequencies are suppressed mildly. The proposed system outperformed an MMSE estimator in subjective listening tests and DRT intelligibility test for MELP-coded noisy speech. The effect of aharmonic comb filtering on the linear predictive coding (LPC) parameters is analyzed using a missing data approach. Suppressing the low-energy frequencies without any modification of the high-energy frequencies is shown to improve the LPC spectrum using the Itakura-Saito distance measure.

The second system combines the aharmonic comb filter with the acoustic-phonetic properties of speech to improve the intelligibility of the MELP-coded noisy speech. Noisy speech signal is segmented into broad level sound classes using a multi-sensor automatic segmentation/classification tool, and each sound class is enhanced differently based on its acoustic-phonetic properties. The proposed system is shown to outperform both the MELPe noise preprocessor and the aharmonic comb filter in intelligibility tests when used in concatenation with the MELP coder.

Since the second noise suppression system uses an automatic segmentation/classification algorithm, exploiting the GEMS signal in an automatic segmentation/classification task is

also addressed using an ASR approach. Current ASR engines can segment and classify speech utterances in a single pass; however, they are sensitive to ambient noise. Features that are extracted from the GEMS signal can be fused with the noisy MFCC features to improve the noise-robustness of the ASR system. In the first phase, a voicing feature is extracted from the clean speech signal and fused with the MFCC features. The actual GEMS signal could not be used in this phase because of insufficient sensor data to train the ASR system. Tests are done using the Aurora2 noisy digits database. The speech-based voicing feature is found to be effective at around 10 dB but, below 10 dB, the effectiveness rapidly drops with decreasing SNR because of the severe distortions in the speech-based features at these SNRs. Hence, a novel system is proposed that treats the MFCC features in a speech frame as missing data if the global SNR is below 10 dB and the speech frame is unvoiced. If the global SNR is above 10 dB of the speech frame is voiced, both MFCC features and voicing feature are used. The proposed system is shown to outperform some of the popular noise-robust techniques at all SNRs.

In the second phase, a new isolated monosyllable database is prepared that contains both speech and GEMS data. ASR experiments conducted for clean speech showed that the GEMS-based feature, when fused with the MFCC features, decreases the performance. The reason for this unexpected result is found to be partly related to some of the GEMS data that is severely noisy. The non-acoustic sensor noise exists in all GEMS data but the severe noise happens rarely. A missing data technique is proposed to alleviate the effects of severely noisy sensor data. The GEMS-based feature is treated as missing data when it is detected to be severely noisy. The combined features are shown to outperform the MFCC features for clean speech when the missing data technique is applied.

CHAPTER I

INTRODUCTION

1.1 Motivation of Research

Despite the significant progress in improving the noise-robustness of speech coders, the intelligibility of parametric speech coders, such as MELPe, still decreases drastically in harsh noise environments [117]. The problem is important particularly for military applications where clearly understanding the correct words is critical. The current NATO standard for military communications is the MELPe speech coder that uses a noise-preprocessor to alleviate the effects of noise. However, the MELPe coder still experiences significant intelligibility degradation in high-noise environments such as in an M2 tank or in a Blackhawk helicopter. Therefore, there is a need for designing speech coders that have high quality and intelligibility in noisy environments.

The speech intelligibility improvement problem is found to be strongly related to the acoustic-phonetic properties of the speech signal. This fact motivated more research in noise-robust speech segmentation using the auxiliary sensors. Therefore, the final part of this dissertation addresses noise-robust automatic speech recognition using the an auxiliary sensor.

1.2 Scope of Thesis

This thesis addresses the intelligibility problem of the parametric speech coders in harsh noise environments by using a noise preprocessor. Since the research is motivated mostly by defense applications, the proposed noise preprocessing systems are concatenated with the MELP speech coder. The MELPe speech coder, concatenation of the MELP coder with a noise preprocessor, is used as the comparison basis. The proposed noise preprocessors use non-acoustic auxiliary sensors, Glottal Electromotor Motion Sensor (GEMS) device,

physiological microphone (p-mic), and bone microphone in addition to the acoustic microphone. The GEMS device is an electromagnetic sensor that can detect the vibrations in the larynx when directed at the throat. P-mic is a gel-suspended microphone that can detect the vibrations on the skin. The bone microphone can receive some of the low frequency speech cues when in contact with the speaker's head.

A Bayesian risk minimization (BRM) framework is proposed that can exploit the perceptual and acoustic-phonetic knowledge of the speech signal in a probabilistic framework. Some heuristic systems that have been shown to outperform the mathematically optimal systems that they are based on are explained theoretically using the proposed framework. Moreover, the proposed framework can be used to develop speech enhancement systems that can optimally incorporate acoustic-phonetic and psychoacoustic knowledge of speech.

The BRM framework is used as a guide to develop two noise suppression algorithms. The first algorithm makes a noise-robust sinusoidal model of noisy speech incorporating the GEMS signal, and uses it for aharmonic comb filtering. The second algorithm uses a multi-sensor segmentation software to automatically segment noisy speech. Acoustic-phonetic knowledge of basic sound classes are then used in addition to the aharmonic comb filter. Both algorithms are shown to improve intelligibility of the MELP-coded speech, when used at the front-end, compared to the MELPe coder.

Noise-robust automatic speech recognition (ASR) using the GEMS signal is investigated to improve the robustness of the segmentation algorithm. A novel system is developed that can efficiently exploit a voicing feature extracted from the GEMS signal. Initial experiments are conducted using a simulated voicing feature extracted from the clean speech signal because of the lack of sufficient sensor data for ASR experiments. Therefore, the additive noise in the GEMS signal, and in most currently available auxiliary sensors, is not considered in the initial experiments. Moreover, the sensor noise can be severe for some subjects that can significantly alter the results. Adverse effects of sensor noise on ASR is investigated. Periodicity and energy features extracted from the GEMS signal are combined with the MFCC features to improve the ASR robustness. A missing data technique is used to alleviate the negative impact of the severe noise that is found in some of the GEMS signals.

1.3 Organization of Thesis

The remainder of this dissertation is organized as follows. A brief background to speech enhancement and noise-robust ASR fields is provided in Chapter 2. A Bayesian risk minimization framework for perceptually-motivated enhancement of the speech signal is presented in Chapter 3. The aharmonic comb filter is described in Chapter 4. In Chapter 5, a noise-robust LPC extraction algorithm is described. The segmentation-based speech enhancement algorithm is presented in Chapter 6. An automatic speech recognition system that efficiently exploits a GEMS-based voicing feature is described in Chapter 7. Effects of severe sensor noise is completely ignored in Chapter 7 because the voicing feature used in the experiments are extracted from the clean speech signal. In Chapter 8, effects of severe sensor noise on the GEMS-based features are investigated by using a database that contains both GEMS and acoustic signals. The dissertation is concluded in Chapter 9 where the contributions of the dissertation are summarized.

CHAPTER II

BACKGROUND

The main contribution of this dissertation is a multi-sensor, segmentation-based noise preprocessor for MELP speech coders. The system has a noise preprocessing aspect, and a noise-robust automatic speech segmentation aspect. Therefore, in this chapter, an overview of the speech enhancement systems is presented in Section 2.1, and noise-robust automatic speech segmentation systems is presented in Section 2.2 to provide a broader view to the problem. Most of the noise-robust automatic segmentation systems borrowed their tools from the noise-robust automatic speech recognition field. Hence, an overview of some of the mainstream noise-robust automatic speech recognition techniques are also discussed in Section 2.2.

The proposed noise suppression systems use the glottal electromagnetic motion sensor (GEMS) device in addition to the acoustic microphone. A description of the GEMS device is presented in Section 2.4. The acoustic microphone and GEMS data that is used for the noise suppression experiments is also described in Section 2.4.

2.1 Speech Enhancement

Speech enhancement can be defined as modifying a distorted speech signal such that the modified signal is as close as possible to the original signal with respect to a subjective and/or objective distance sense.

Speech can be distorted by many different sources. Two of the most commonly studied distortion sources are the additive ambient noise and reverberation. In this research, only additive ambient noise is considered.

The measures that are used to assess the performance of a speech enhancement system depend on the goal of the system. For example, a perceptual quality improvement task and an intelligibility improvement task require different measures to assess the performance.

Commonly used subjective and objective measures are discussed in Section 2.1.5.

The majority of the work in the speech enhancement field is done using single microphone signals. Recorded noisy signal is typically enhanced using statistical models for noise, clean speech, and noisy speech signals. Some of the more recent enhancement techniques, such as subspace methods, are reviewed in [39], [38]. The paper in [83] discusses some of the earlier systems such as spectral subtraction, Wiener filtering, adaptive comb filtering, and Linear Predictive Coding (LPC) based speech enhancement. Some of the more popular, single microphone speech enhancement systems are described below.

2.1.1 Single Microphone Statistical Speech Enhancement

2.1.1.1 Spectral Subtraction and Wiener Filter

Spectral subtraction is one of the earliest speech enhancement techniques. Statistically independent random variables are used to model both speech and noise. Power spectrum of clean speech is estimated with a maximum likelihood (ML) technique.

Spectral subtraction is effective in removing noise from noisy speech. However, it usually leaves an annoying residual noise, the musical noise, in the enhanced signal. There are many variations of spectral subtraction that address the musical noise problem [102], [87], [51]. A detailed analysis of the spectral subtraction technique is given in Section 3.2 where it is discussed in the context of a Bayesian risk minimization framework.

Using the same statistical models, the problem of power spectrum estimation can also be solved using a minimum mean square error (MMSE) estimator. This method of speech enhancement is called Wiener filtering [83].

2.1.1.2 MMSE Amplitude Estimator

Spectral subtraction and Wiener filtering techniques approach the problem from a power estimation perspective. However, the ultimate goal of enhancement is, typically, estimating the spectral amplitude. An MMSE estimator for short-time spectral amplitude estimation is proposed in [40]. Another MMSE estimator that minimizes the log-spectral amplitude distortion is proposed in [41].

An MMSE estimator with a gamma-distribution assumption, as opposed to the Gaussian

distribution assumption in [40], is proposed in [90]. Computationally efficient approximations of the MMSE estimator are proposed in [143].

The MMSE estimator in [40] is famous for its capacity to eliminate the musical noise. In [9], a detailed analysis is done to show the reasons why musical noise does not occur in [40]. Motivations for using an MMSE estimator for speech enhancement are discussed in [35] where the MMSE estimator is shown to be optimal for some of the distance measures that are commonly used in speech processing.

2.1.1.3 Signal-to-Noise Ratio (SNR)-Adaptive Systems

Most speech enhancement techniques use SNR in a statistical estimation framework. However, those systems that are mathematically optimum do not always generate perceptually high quality speech. One of the approaches to this problem is using an adaptive system that modifies the optimum noise suppression factor based on the estimated SNR. Typically, these systems use aggressive suppression at lower SNRs and mild suppression at higher SNRs. Although the final system is typically mathematically suboptimum, it can successfully handle the musical noise artifact.

An adaptive Wiener filter is proposed in [142] where the SNR estimates are smoothed adaptively between speech frames in the stationary speech and noise sections to avoid the artifacts common to Wiener filter.

2.1.1.4 Subspace Methods

Signal subspace methods decompose the noisy speech signal into a noise plus speech component and a noise-only component. The noise-only component is suppressed without distorting the speech signal in [43]. An extension to the colored noise case is proposed in [82].

An interesting aspect of the signal subspace methods is their noise shaping capability. These systems can conveniently leave some amount of noise with the speech that is shaped to be tolerable by the listener while significantly preserving the original speech signal. In [68], a spectral domain signal decomposition method is proposed that shapes the residual noise based-on the masking threshold of the human auditory system. The goal is to reduce the

musical noise artifact that is common in the subspace methods by adjusting the attenuation factor of the suppression filter based on the energy of the speech signal in a critical band.

In [66], a prewhitening method is used so that the subspace method can be applied to any type of colored noise. The system in [66] is also shown to be a generalization of the system in [43] which was designed for white noise.

Wavelets can also be used to decompose the noisy speech signal into noisy speech and noise-only components. Typically, wavelet coefficients are hard-thresholded to separate noise and speech signal. However, such a hard-threshold separation can cause distortion to the speech signal. In [6], soft-thresholding is applied that changes based on the masking level at each subband.

2.1.1.5 Signal Presence Uncertainty

One of the flaws of the earlier statistical methods is the assumption of a speech plus noise model at all points in the time-frequency plane. However, speech signal does not necessarily appear at all times and frequencies in a typical speech recording. An obvious example to that is the silence periods in the signal. Moreover, even when the speech signal exists, some frequencies may contain negligible amount of speech energy. For example, voiced sounds contain considerably higher energy at their harmonic frequencies compared to the non-harmonic frequencies. Thus, a soft-decision method can be used to detect the existence of speech at a particular time-frequency location, and the enhancement system can take into account the signal presence uncertainty. The system proposed in [15], for example, modifies the MMSE estimator proposed in [41] to include the signal presence uncertainty. A global soft-decision method is proposed in [75] where each speech frame is assigned a single signal presence uncertainty (P_s) value. Hence, it can address the pauses during speech. A similar two-state model is proposed in [97].

The global soft-decision method does not take into account the cases where speech signal exists only at certain frequencies in a time frame. The system proposed in [89] assigns a different P_s value to each time-frequency location. The idea is to consider sounds like vowels where the speech signal is quasi-stationary and most of the energy is concentrated at the

harmonic locations.

Hard-decision and soft-decision adaptive speech presence detection algorithms are proposed in [127]. Similar to [89], different P_s values are assigned to different time frequency locations.

2.1.1.6 *Enhancement with Hidden Markov Models (HMMs)*

Most of the earlier statistical speech enhancement systems ignore the fact that different speech sounds have different acoustic characteristics that can be used to improve the performance of a speech enhancement system. HMM-based enhancement systems attempt to exploit some of the structure in the speech signal. A detailed review of the HMM-based enhancement systems is presented in [37].

HMM models for both noise and speech signals are used to derive MMSE and maximum a posteriori (MAP) estimators for speech enhancement in [36]. Autoregressive (AR) parameters are used for building HMMs, and Wiener filtering is used to enhance speech at each HMM state. HMM-based systems are iterative with the idea that more accurate state sequences can be obtained at each iteration. In [88], the noise model is also updated at each iteration.

HMM-based enhancement techniques use the Expectation Maximization (EM) algorithm which is sensitive to initialization of the models. However, especially with the severely noisy signal, initialization points are typically poor which leads to saturation at a local maxima in the search space. Initialization with a noise compensated AR-HMM is proposed in [125]. HMM models for speech and noise are combined using a model combination technique in the cepstral domain to obtain models for noisy speech, and the noisy speech models are used for initialization.

Mel Frequency Cepstrum Coefficients (MFCC) are known to perform better than the AR parameters in ASR systems. In [95], an MFCC-based system is used to initialize the system. After the alignment of states with the noisy signal, AR parameters of each state are used to find the optimum enhancement filters. Clean HMM-models are adapted to unknown noise environment using data-driven parallel model combination (DDPMC).

One of the problems with the HMM-based methods is its computational complexity. A low complexity method is proposed in [81] where only the current and next states are considered for state transitions as opposed to all states to all states connections in the other systems. In [122], a system with real-time implementation considerations is proposed.

2.1.1.7 Model-based Speech Enhancement

There are speech enhancement systems that use the mathematical models of the speech signal. Two of the more popular models are the LPC model and the sinusoidal model. Examples of enhancement systems using these two models are discussed in this section.

In [23], enhancement is performed on the LPC residual signal to obtain sharper harmonics in voiced speech, and the LPC spectrum is driven by the enhanced residual signal. LPC residual is also used to make noise/speech decision in [23].

A constrained iterative speech enhancement technique is described in [69]. The idea is to keep the smoothness constraints of the enhanced speech signal to avoid some of the possible artifacts introduced by the enhancement algorithm.

In [104], higher order statistics of the speech signal is used to distinguish speech from additive Gaussian noise. Speech spectrum is divided into narrow bands where noise spectrum is assumed to be flat in each band. Fourth order cumulant of the noisy speech statistics is used to estimate sinusoidal speech parameters in each band.

2.1.1.8 Perceptually-Motivated Enhancement Systems

Perceptual aspects of the human auditory system should be taken into account in quality and intelligibility improvement tasks. Statistical models that do not consider human perception can generate audible artifacts. Hence, statistically minimizing the residual noise energy in the enhanced speech signal does not always generate high quality speech. Several systems that consider various aspects of the human auditory system is discussed in this section.

In [118], an adaptive Wiener filter is proposed that takes into account the sensitivity of the human ear to the rapid spectral changes in the speech signal. Such changes occur, for example, at the vowel transitions and plosive bursts. Human ear can successfully mask the

additive noise at those instances. Therefore, the Wiener filter is adapted such that it preserves the rapid changes in the nonstationary regions of the speech signal while suppressing the additive noise. The adverse effects of excessive smoothing and loss of important speech cues are avoided in this method.

In [136], critical-band masking properties of the inner ear is considered in the enhancement. The spectral subtraction parameters are tuned such that some amount of residual noise that can be masked by the speech signal is left in the enhanced speech. The residual noise allows flexibility in enhancement since it is not audible.

In [78], a perceptual post-processing filter is employed after a signal subspace enhancement scheme to reshape the residual noise. Two criteria are considered. Noise peaks that the ear is sensitive to are smoothed out, and residual noise is suppressed to be below the masking threshold.

In [12], spectral masking property of the human ear is used for enhancement. A strong tone can mask a weaker tone with higher chance if the weaker tone is spectrally closer to the strong tone. Therefore, the noise at the low-energy spectral valleys are severely suppressed while the lateral inhibition property of the ear is used to suppress the noise at the spectral peaks.

2.1.2 Multi-Sensor Speech Enhancement

Multi-sensor speech enhancement systems use non-acoustic auxiliary sensors in addition to the acoustic microphone to effectively enhance the noisy speech.¹ Some of the recently proposed multi-sensor speech enhancement systems are discussed below.

In [65], a glottal correlation filter is proposed that uses the General Electromagnetic Motion Sensor (GEMS) device [117]. Significant intelligibility improvement over the MELPe coder using the Diagnostic Rhyme Test (DRT) is reported.

The system proposed in [117] fuses the GEMS signal, physiological microphone (p-mic) signal, and the bone microphone signal with the acoustic microphone signal based on the estimated quality of those signals in each frequency subband.

¹Microphone array based systems and multiple microphone systems are also sometimes called multi-sensor systems, but they are out of the scope of this section.

The system proposed in [149] uses the bone-conduction microphone in addition to the acoustic microphone. Speech/non-speech detection is done using the bone-conduction microphone for improved speech quality, and noisy speech signal is enhanced using the bone-conduction microphone signal in a Wiener filtering framework. The automatic speech recognition (ASR) results for the system in [149] are presented in [147].

2.1.3 Noise Estimation

Noise estimation is a fundamental block in most of the speech enhancement systems. Therefore, significant amount of research has been done on noise estimation and effects of additive noise on the speech signal [106], [150].

One of the more recent and popular noise estimation techniques uses an optimum smoothing technique to estimate the noise power spectral density (psd) using minimum statistics [90]. The idea is to multiply the minimum statistics for each time-frequency location by a bias factor using optimum smoothing. The system proposed in [14] also uses the minimum statistics. However, a simple recursive averaging technique is used in [14] based on signal presence probability estimation.

2.1.4 SNR Estimation

SNR is defined as

$$SNR(n, k) = 10 \log_{10} \left[\frac{S(n, k)}{N(n, k)} \right] \quad (1)$$

where $S(n, k)$ represents the power of the speech signal and $N(n, k)$ represents the power of the noise signal at time n and frequency k . Estimated SNR is commonly used as an input to the statistical speech enhancement systems. Hence, SNR estimation is an important block in most speech enhancement systems.

A non-casual SNR estimation method is proposed in [17] where not only the past values of the noisy speech samples, but also the future values are taken into account for estimating SNR. The popular “decision-directed” SNR estimation method, which uses the past estimations to estimate the current SNR, is analyzed as a special case of the noncausal estimator in [16].

In [131], the input signal is divided into 32 auditory channels, and SNR in each band is estimated using neurophysiologically-motivated spectro-temporal features. The idea is to extract the features by mimicing the ear and use those features in a neural network to estimate SNR.

2.1.5 Assessing the Performance of a Speech Enhancement System

The method for assessing the performance of a speech enhancement system depends on the goal of the system. An enhancement system can be designed for perceptual quality improvement, intelligibility improvement, or as a preprocessor to an ASR system. Subjective listening tests, such as A/B test or MOS test, are used for quality improvement tasks [116]. For intelligibility tasks, diagnostic rhyme test (DRT) or similar methods are used [116]. Enhancement for ASR tasks are typically assessed with ASR experiments.

Besides the subjective measures, there are also objective distance measures as described in [54] and [116]. Objective distance measures are used to mathematically define the distortion between the enhanced speech and original speech. They have been used to predict the performance of a speech enhancement system for both perceptual quality improvement and ASR preprocessing tasks with limited success.

One of the more commonly used objective measures is the segmental-SNR (S-SNR) measure [116]. The speech signal is divided into short-time signal frames. Segmental-SNR for each noisy speech frame i is

$$S - SNR_i = 10 \log_{10} \left(\frac{S_i}{N_i} \right). \quad (2)$$

where S_i is the signal power, and N_i is the noise power in frame i . Average segmental-SNR of N frames is

$$S - SNR = \sum_{i=1}^N S - SNR_i \quad (3)$$

Itakura-Saito distance (d_{IS}) is another commonly used distance measure [116]. For a clean speech spectrum $S(w)$ and a distorted speech spectrum $S'(w)$

$$d_{IS} = \int_{\pi}^{-\pi} \left[e^{V(w)} - V(w) - 1 \right] \frac{dw}{2\pi} \quad (4)$$

where

$$V(w) = \log[S(w)] - \log[S'(w)]. \quad (5)$$

$V(w)$ is called the log-likelihood ratio (LLR) and sometimes used as a distance measure (d_{LLR}) [116]. If the vector of linear prediction (LP) coefficients for the clean speech signal is a_c , the vector of LP coefficients for the distorted signal is a_d , and the autocorrelation matrix for the clean speech signal is R_c , then

$$d_{LLR} = \log \left(\frac{a_d R_c a_d}{a_c R_c a_c} \right). \quad (6)$$

2.2 Noise-Robust Automatic Speech Recognition

State-of-the-art Automatic Speech Recognition (ASR) systems can achieve less than point two percent word error rate for the digit recognition task. Moreover, they have high performance for medium and large vocabulary dictation tasks. However, these performance results are obtained in quiet environments. Performance is known to drop drastically for real-life environments where speech is contaminated with an unpredictable noise source resulting in an unknown and time-varying signal-to-noise Ratio (SNR). Therefore, noise robustness is one of the key challenges remaining to be solved before the ASR systems can be employed in real environments.

Current approaches to the noise robustness problem can be classified into several broad categories. Some techniques compensate for the effect of noise on speech features and create a better match between acoustic models and noisy features [80], [52]. Similarly, acoustic models can be modified to better represent the features extracted in a noisy environment. Parallel Model Combination [48] and model adaptation approaches [50] [49] are in this category.

In this dissertation, missing data techniques, noise-robust features, and speech enhancement are used for noise-robust automatic segmentation. Therefore, these three approaches are discussed in more detail below.

2.2.1 Missing Data Techniques

The missing data techniques approach the noise robustness problem from a reliability perspective. In contrast to the methods mentioned above, missing-data based systems do not compensate the features or the acoustic models. Instead, they attempt to detect the reliable features and do recognition based only on those reliable features. Thus, the two challenges these methods face are detecting the unreliable (missing) features and decoding the noisy speech by using the reliable features.

Spectral subtraction and SNR thresholding are two popular ways of estimating the subband reliability [20]. Once the reliable bands are detected, either marginalization or data imputation approaches can be used for recognition. In the marginalization technique, the likelihood function of the reliable features are estimated through integration over unreliable features. In the data imputation technique, the missing features are imputed from the reliable features.

Early work on missing data techniques include [77], [63], [135]. These systems were based on detecting of unreliable parts of the spectrum in a hard-decision manner and imputing the unreliable parts using the estimated noise level. Although most of these systems use subband features, a cepstral domain version has also been proposed in [99].

Recently, a theoretical framework for handling missing data using the marginalization technique has been presented in [20]. Performance of the system is assessed for both small vocabulary and large vocabulary tasks [10]. Recent systems that use data-imputation algorithm in the cepstral domain include [56], [10], [120].

2.2.2 Noise-Robust Feature Extraction

A strategy for improving noise robustness of ASR systems is using noise-robust features. A review of the popular noise-robust feature extraction schemes is discussed in [114]. Some of the front-end signal processing techniques used for noise-robust ASR is described in [67]. Examples of noise-robust features are discussed below.

2.2.2.1 Features Derived from DFT Spectrum

Mel-Frequency cepstrum coefficients (MFCC) are the most popular features in ASR systems [114]. However, cepstrum-based features are sensitive to noise [108]. Therefore, alternative features have been considered that are directly derived from the power spectrum.

In [109], filter-bank energies are fed into the recognizer without further processing. It is shown that filter-bank energies perform at least as good as the MFCCs in noisy environments. In [145], difference between the power spectrums of the consecutive frames are used to derive the MFCC features. The idea is to eliminate the background noise that is assumed to be stationary over the subtraction interval. A similar strategy is described in more detail in [11]. In [123], instead of the logarithmic nonlinearity in the MFCC features, an n -th order root operation is performed to improve noise robustness. These features are called root-cepstrum coefficients (RCCs).

Two dimensional (2D) cepstrum coefficients represent a matrix of feature vectors that can efficiently incorporate the temporal dependence between the feature vectors [13]. In [85], the 2D cepstrum coefficients are modified using genetic algorithms to obtain noise robustness.

2.2.2.2 Features Derived from LPC Spectrum

MFCC features are derived from the Discrete Fourier Transform (DFT) spectrum of the speech signal. However, LPC spectrum can also be used to extract the features.

DFT-based features are known to outperform the LPC-based features in clean conditions[94]. Still, because of their parametric nature, LPC-based features can be useful for noise-robust feature extraction. For example, in [61], one-sided autocorrelation sequence is used for linear prediction of noisy speech which is then used to extract noise-robust LPC features.

2.2.2.3 Perceptually-Motivated Features

Human auditory system is significantly more robust to noise than the current ASR systems. Some of the perceptually-motivated feature extraction methods that mimic the human ear are discussed below.

In [146], a minimum variance distortionless response (MVDR) of the speech spectrum is used to extract noise-robust features. It is shown that the MVDR method can accurately track the peaks in the spectrum while being computationally simple. Moreover, recognition accuracy improves significantly in noisy conditions compared to the popular Perceptual Linear Prediction (PLP) features.

RASTA is another perceptually-motivated, noise-robust feature extraction method [60]. The idea is to filter out the spectral components of speech that change faster or slower than the natural range of change rate of the speech signal. Significant reduction of noise can be obtained with this method.

Speech signal has resonant frequencies that are important determinants of the perceptual speech quality. In [133], formant locations as well as some perceptually distinctive acoustic features are combined with the MFCC features in a multi-stream recognition framework. Significant decrease in WER is reported in noisy conditions.

Filterbanks are used to mimic the human auditory system in the MFCC features. However, the bandwidths of these filters are typically chosen to be triangular while it has arbitrary shapes in the auditory systems. More precise modeling of the filter bandwidths is shown to improve noise robustness in [126].

2.2.2.4 Environment-Dependent Feature Selection

Different features can have higher performance in different background environments. In [2], an environmental sniffing algorithm is proposed that measures the characteristics of the environmental noise. The features that are expected to perform best for the detected environment are used to do the recognition.

2.2.2.5 Aurora Evaluations

Aurora, a working group of European Telecommunications Standards Institute (ETSI), has been found to address some of the issues related to distributed speech recognition. In this context, a standard speech recognition setup is prepared to measure the performance of noise-robust front-ends given all the other parameters of the recognizer are same at all research sites. Evaluations are done using the standard Aurora speech databases. Results of

these evaluations with many comparison papers can be found in the International Conference on Speech and Language Processing (ICSLP) 2002 conference papers and in some of the later conferences [21].

2.2.3 Speech Enhancement for Noise-Robust ASR

Speech enhancement is a commonly used method to reduce noise in the speech signal before feeding it into an ASR system. There are numerous papers that compare the performance of various speech enhancement systems on automatic speech recognition. In [57], performance of spectral subtraction, nonlinear spectral subtraction [87], MMSE algorithm [40], and noise adaptive AutoLSP algorithms [69] are compared using the Aurora2 corpus. The nonlinear spectral subtraction method is found to be the best performer among all the candidates. Another study with nonlinear spectral subtraction using Aurora 2/3 tasks is presented in [45].

An interoperability study of speech enhancement and ASR systems is done in [103]. It is shown that enhancement systems improve the performance compared to the baseline system at lower SNRs. However, if the ambient noise is mild, speech enhancement can have an inverse effect particular generating deletion errors.

2.2.4 Feature Enhancement

ASR systems use features extracted from the speech signal for recognition. Therefore, estimating the clean features from the noisy features may be more suitable for an ASR task as opposed to speech enhancement that attempts to estimate the clean speech.

In [53], a probabilistic optimum filtering framework [105] is used to estimate the clean speech features from the noisy speech features using a throat microphone. The idea is to divide the clean feature space into vector quantized (VQ) regions. The noisy input feature is mapped to one of those VQ regions in a soft-decision manner. The throat microphone signal is used to improve the mapping vector.

Noisy features can be modeled with a probability distribution function (pdf) such as a Gaussian. The uncertainty in the noisy features can be handled with a statistical uncertainty decoder in this case. For example, in [32], uncertainty is addressed in the feature space by

using the estimated variances of the enhanced features.

2.3 Noise-Robust Speech Segmentation and Classification

Segmentation can be defined as dividing a speech signal into segments that contain one and only one predefined sound class. For example, a speech signal can be segmented into broad sound classes such as vowels, nasals, stops, etc.

Although segmentation can automatically generate the segments, it does not classify the segments. Classification can be defined as labeling the segments that are generated by the segmentation algorithm. HMM-based systems perform segmentation and classification simultaneously. However, there are systems that perform these two tasks separately as discussed below.

2.3.1 Discriminative Feature Extraction

One of the biggest challenges in an automatic segmentation task is finding the features that are suitable for discriminating between the sound classes. In [134], narrowband and wideband classifiers are used to detect segment boundaries that correspond to narrowband and wideband diphone boundaries. Boundaries that are detected independently are combined together using lowest-cost path algorithms.

In [18], duration of phonemes are used as a confidence measure for segmentation. Although the method is independent of the segmentation paradigm, the results are reported for HMM-based systems in noisy conditions. Similar to [18], phoneme duration is used in conjunction with the acoustic similarity measures using a genetic algorithm in [129].

A soft-decision hierarchical method using support-vector machines is proposed for automatic segmentation in [70]. Sound classes become more specific down the hierarchy with a speech/non-speech detection at the first branch. Both segmentation and classification are done for recognizing broad sound classes.

2.3.2 HMM-based Techniques

Similar to ASR systems, hidden Markov models (HMMs) are commonly used for automatic segmentation. An overview of HMM-based approaches to automatic speech segmentation

is presented in [31] where Viterbi decoding is compared with forward-backward decoding.

Although HMM-based techniques typically perform well in detecting segments, segment boundaries are often not accurate. Therefore, spectral boundary correction is commonly used for HMM-based segmentation systems [76] Postprocessing techniques for fine-tuning the segment boundaries in HMM-based systems are discussed in [31].

In [115], broad-level sound classification is done using HMM methods. Features that are effective in making classifications between two sound classes are used in a hierarchical manner. Some of the popular noise-robust ASR techniques are applied to noise-robust segmentation and compared with each other in [111].

2.3.3 Spectral Variation-based Techniques for Segmentation

Automatic segmentation typically generates segments that have common spectral properties. Hence, spectral-variation can be a good measure to detect the segment boundaries. Broad-level speech segmentation is done in [112] for spontaneous speech. Three different feature sets are used that are compared for speech contaminated with white noise. Various features, similarity measures, and liftering algorithms are compared in [121] for detecting the transition regions in speech.

2.4 Description of the GEMS Device and the ARCON Database

The glottal electromagnetic motion sensor (GEMS) is a micro-power device that can be used, among other things, to detect motion in larynx. The GEMS device consists of a penetrating radar whose principles have been studied both at the Lawrence-Livermore Laboratory and Aliph, Inc. A fully developed, commercial version based on these principles is currently available from Aliph, Inc. Descriptions of its properties can be found in [8].

When positioned correctly on the exterior of the throat adjacent to the larynx, the output of the radar during voiced speech is a signal that resembles the airflow volume velocity at the larynx. The exact physical structures whose motion are detected are currently not completely understood. The signal, however, is often stable and as such can be useful in further processing. Additionally, the signal obtained is robust to external acoustic influences,

such as ambient noise.

The GEMS device responds to vocal fold vibration at the larynx. Other devices such as EGG do this by measuring the changes in conductivity at the throat, but it is considered too cumbersome for everyday use. The GEMS device can be deployed in handheld devices for speech applications, such as noise robust pitch detection [8] and speech enhancement [7], in a less intrusive manner.

2.4.1 ARCON Database Description

The GEMS system was investigated for very low bit-rate speech coding (300 bps) in noisy environments as part of a Department of Advanced Research Projects Agency (DARPA) project [7]. As part of this program, an extensive database was created by ARCON Corporation having simultaneous speech, GEMS, EGG, and other sensor data for various military noise conditions [119]. Recordings from ten male and ten female speakers are available in the database. The data is designed for the assessment of speech coding systems. DRT (diagnostic rhyme test) and CVC (consonant vowel consonant) word lists are provided for intelligibility assessment. Harvard sentences and isolated vowels are provided for perceptual quality assessment. The systems proposed in this thesis are tested using the ARCON database. DRT sequences are used for intelligibility tests, and Harvard sentences are used for quality tests.

2.4.2 Correlation of GEMS-based and Acoustic Microphone-based Features

The periodicity feature, the energy feature, and the high-energy frequencies in the GEMS spectrum are found to be strongly correlated with the same features extracted from the clean speech signal for voiced speech. Descriptions and discussions of the three features are done below.

2.4.2.1 Periodicity Feature

The periodicity feature for a short-time GEMS signal frame i (g_i) is defined as

$$v_i = \arg \max_{2.5 \text{ msec} < \tau < 10 \text{ msec}} R_{gg}(\tau) \quad (7)$$

where $R_{gg}(\tau_i)$ is the normalized autocorrelation function of g_i .

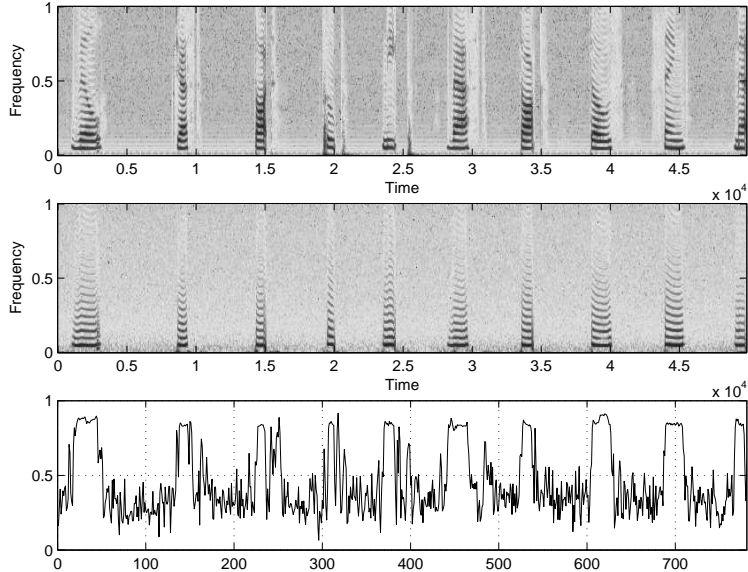


Figure 1: Spectrogram of a sample speech signal is shown in the top figure, and spectrogram of the simultaneously recorded GEMS signal is shown in the middle figure. Trajectory of the periodicity features is shown in the bottom figure. Speech and GEMS samples are taken from the DRT sequences in the ARCON database recorded in the quiet conditions by a female speaker. The parameters given in Table 1 are used in the analysis. The sequence of monosyllables in the utterance are: vill, jest, taught, coop, neap, vast, dock, those, sing, and met.

The GEMS signal is strongly correlated with the airflow velocity at the larynx [8] and indicates high degree of periodicity when the speech signal is voiced as shown in Fig. 1. Moreover, assuming short-time periodicity, the period within a short-time ($\sim 20msec$) frame is equal to the pitch period of the speech signal recorded simultaneously.

2.4.2.2 Energy Feature

The energy feature

$$E_i = \sum_{k=1}^N (g_i(k))^2 \quad (8)$$

is defined as the total energy of the short-time sensor signal g at time frame i . The energy trajectory of the speech signal is highly correlated with the energy trajectory of the GEMS signal as shown in Fig. 2. Moreover, the energy feature is found to be smoother than the voicing feature in most cases. This can be observed when Fig. 1 and Fig. 2 are compared.

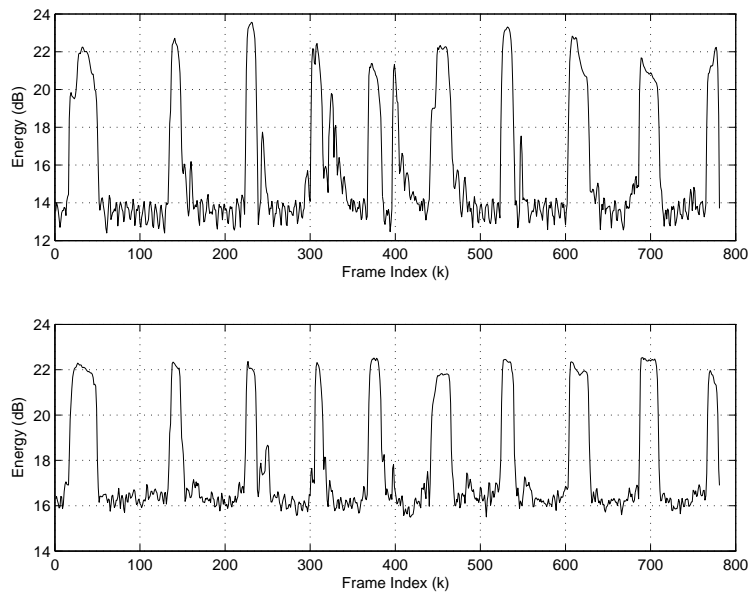


Figure 2: Trajectory of the energy feature is shown for the speech signal in the top figure. Trajectory of the energy feature is shown for the GEMS signal in the bottom figure. Speech and GEMS samples are taken from the DRT sequences in the ARCON database recorded in the quiet conditions by a female speaker. The parameters given in Table 1 are used in the analysis. The sequence of monosyllables in the utterance are: vill, jest, taught, coop, neap, vast, dock, those, sing, and met.

Table 1: The parameters that are used in analysis of speech and GEMS signals in this chapter.

| Attribute | Value |
|---------------|-------------------|
| Window Length | 256 samples |
| Window Type | Hanning window |
| Frame Rate | 128 samples/frame |
| Sampling Rate | 8 kHz |

2.4.2.3 High-Energy Frequency Locations

The GEMS signal often has an harmonic structure up to 4 kHz as shown in Fig. 3. Although the higher frequency harmonics in the GEMS spectrum, most of the time, are not visible in the spectrogram because of their relatively low energy, the spectral peaks are clearly visible when the GEMS spectrum for a short-time frame is inspected as shown in Fig. 3.

The high-energy frequencies in the GEMS and speech spectrums are also found to match most of the time for voiced speech which can be useful in many applications. For example, an aharmonic comb filtering method is described in Chapter 4 that uses the high-energy frequencies in the short-time GEMS frames for further enhancing the noisy audio signal.

2.4.3 Time Delay Between the Acoustic Sensor and the GEMS Sensor

There is a small ($\sim 2msec$) time difference between the GEMS signal and the acoustic signal when they are recorded simultaneously. The asynchrony is a result of the instrumentation delay and the difference in the acoustic wave velocity and the electromagnetic wave velocity. Moreover, the GEMS device measures the airflow velocity at the larynx, but the acoustic microphone measures the airflow pressure close to lips. Therefore, the differences in places of measurement also adds to the delay.

A MATLAB function, developed at Georgia Tech, is used to synchronize GEMS and speech signals. The function is based on a crosscorrelation method. The optimum delay is found at the lag where the cross-correlation of the residual of the acoustic signal and the derivative of the GEMS signal makes a peak.

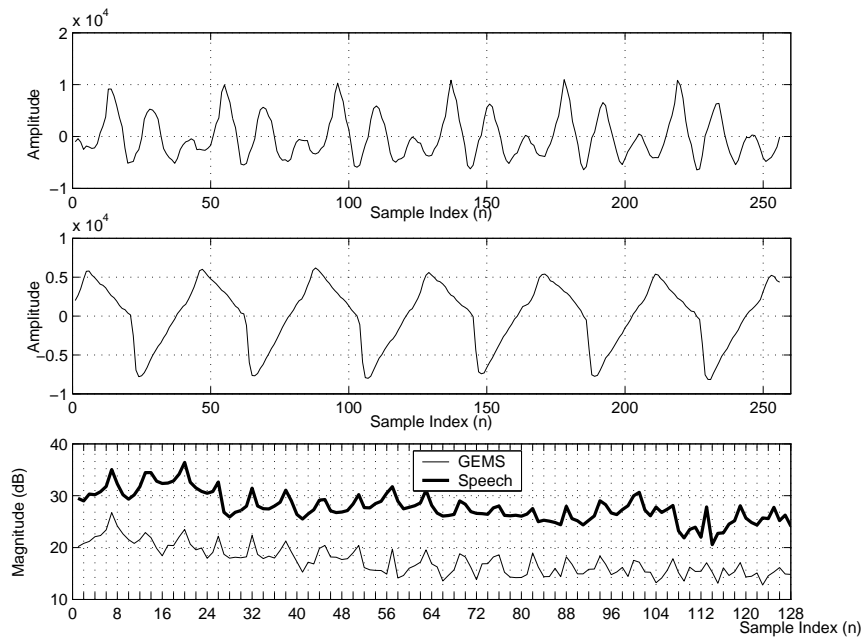


Figure 3: Magnitude spectrums of the GEMS signal and acoustic signal are compared for a voiced speech frame. The time-domain speech signal is shown in the top figure, the time-domain GEMS signal is shown in the middle figure, and the spectrums of both signals are compared in the bottom figure.

CHAPTER III

A PERCEPTUALLY-MOTIVATED BAYESIAN RISK MINIMIZATION FRAMEWORK FOR SPEECH ENHANCEMENT

A typical noise suppression system always carry a risk of misestimating the original signal, and perfect reconstruction is often not possible. These misestimations are associated with perceptual costs that depend on the goal of the system. To minimize the average cost of misestimations, a statistical speech enhancement framework should have the following properties

1. Sources of misestimations have to be identified, and perceptual costs of possible misestimations should be explicitly taken into account.
2. Speech signal is nonstationary, and different speech sounds have different acoustic and perceptual properties. Therefore, in order to accomplish the first requirement, the framework should be able to incorporate acoustic-phonetic knowledge of speech and perceptual aspects of the human auditory system.

Traditional systems such as Wiener filters, spectral subtraction filters, and MMSE estimators do not explicitly consider the costs associated with possible misestimations. The deficiency of the traditional systems in addresssing the above criteria have been investigated, and, typically, optimum linear enhancement filters are nonlinearly modified to improve the perceptual quality of the enhanced speech signal. Most of those nonlinear modifications are based on observations and heuristic algorithms.

The heuristic algorithms typically undersuppress or oversuppress the noisy signal to preserve the speech cues and suppress the noise. Oversuppression is achieved by subtracting more noise than dictated by the optimal filter. Similarly, undersuppression is achieved by

subtracting less noise than dictated by the optimal filter. For example, there are many nonlinearly modified spectral subtraction filters that use SNR-based oversuppression factors and perform better than the mathematically optimum filter [79], [102], [87], [51]. There are also systems that take into account the perceptual properties of the human auditory system by post-processing the enhanced speech signal. Masking property of human perception, for example, is commonly used in such systems [12], [78].

Signal presence detection framework is another example of incorporating an SNR-based oversuppression factor into the enhancement framework [15], [41], [75]. The suppression factor is adapted based on the probability of existence of the speech signal in the noisy signal. Noisy signal is more aggressively suppressed with decreasing probability of speech existence. SNR-based measures are typically used to estimate the probability of existence.

The signal presence detection framework does not take into account the uncertainty in the SNR parameter. Low SNR does not always mean low probability of speech existence and high SNR does not always indicate signal presence partly because of possible SNR misestimations. Moreover, SNR-based signal presence detection does not consider the perceptual aspects of the human auditory system and the acoustic-phonetic properties of the speech signal.

In the next section, a statistical speech enhancement framework is described that is based on Bayesian risk minimization.

3.1 Proposed Framework

In the proposed framework, the speech enhancement filter is designed to minimize the costs of misestimations. Using the risk and the perceptual costs of misestimations, the perceptually optimum estimator can be mathematically derived. Several assumptions have been made to simplify the formulation:

1. System is assumed to be causal. Therefore, effects of future values are not considered in calculating the risk function, and post-processing of the enhanced speech is not considered.
2. A subband-based speech enhancement method is used where each band represents an

auditory channel. Auditory channels are assumed to be independent, and misestimations in a band do not have a cost in the other bands.

3. The system is sound-class dependent. To exploit the acoustic-phonetic properties of speech sounds, speech is segmented into broad level sound classes. One particular sound classification is proposed in Chapter 6.
4. The cost function is related to the sound class of the current frame only, and sound classes of the earlier frames are not taken into account. Transition regions between speech sounds can be included as another class of sound to avoid problems in these perceptually important regions.

The variables that are used to formulate the risk function are defined as follows. The gain factor that perfectly recovers the original signal is denoted by G while the gain factor estimated by the speech enhancement system is denoted by \widehat{G} . Similarly, the vector of previous perfect gain factors at time i and auditory channel j is

$$\underline{G}_{i,j} = [G_{(i-1),j} \ G_{(i-2),j} \ \dots \ G_{2,j} \ G_{1,j}]^T \quad (9)$$

while the vector of previously estimated gain factors is

$$\underline{\widehat{G}}_{i,j} = [\widehat{G}_{(i-1),j} \ \widehat{G}_{(i-2),j} \ \dots \ \widehat{G}_{2,j} \ \widehat{G}_{1,j}]^T. \quad (10)$$

$G_{i,j}$ and $\widehat{G}_{i,j}$ denote scalar gain factors obtained at time i and auditory channel j . Previous values of the clean signal at time i and auditory channel j are

$$\underline{X}_{i,j} = [X_{(i-1),j} \ X_{(i-2),j} \ \dots \ X_{2,j} \ X_{1,j}]^T \quad (11)$$

while the vector of previously estimated speech values is

$$\underline{\widehat{X}}_{i,j} = [\widehat{X}_{(i-1),j} \ \widehat{X}_{(i-2),j} \ \dots \ \widehat{X}_{2,j} \ \widehat{X}_{1,j}]^T. \quad (12)$$

For a sound class c , the average perceptual cost (C) of using the gain factor $\widehat{G}_{i,j}$, risk function, is

$$E[C(\widehat{G}_{i,j})] = \int_{-\infty}^{\infty} C(\widehat{G}_{i,j}/\underline{\widehat{X}}_{i,j}, c, G_{i,j}) P(G_{i,j}/\underline{\widehat{X}}_{i,j}, \widehat{G}_{i,j}, c) d(G_{i,j}) \quad (13)$$

where c denotes the sound class.

In Eq. 13, acoustic-phonetic properties of the speech signal are taken into account through formulating the problem in terms of the sound class c . Probability of misestimation (P_m) is incorporated with $P(G_{i,j}/\hat{X}_{i,j}, \hat{G}_{i,j}, c, G_{i,j})$ that represents the probability of a suppression factor $G_{i,j}$ given the previous estimates of the signal $\hat{X}_{i,j}$, previous suppression gains $\hat{G}_{i,j}$, and the sound class c .

Perceptual knowledge of the human auditory system can be included into the estimation process through the cost function C . For example, in the musical noise case, if there are narrow-band islands left after enhancement, even if those regions have high SNR and low risk values, they can be suppressed by the system by assigning high cost to such situations. Some of the typical sources of misestimations and the perceptual costs of misestimation are discussed below.

3.1.1 Sources of Misestimations in Speech Enhancement Filters

In this work, two sources of misestimations are identified. The first source is the variance of the estimator for the given input parameters. In the case of speech enhancement, the uncertainty in the estimated value, and therefore P_m , rapidly increases with decreasing SNR. The second source of misestimation is the uncertainty in the input parameters. For example, SNR misestimations can significantly increase P_m in speech enhancement. Robustness of an estimator to the uncertainty in the prior parameter is studied in the Robust Bayesian Analysis (RBA) field [79]. The methods that can be used to quantify the uncertainty in the apriori parameters are discussed in more detail in Section 3.1.3.

3.1.2 Costs of Misestimations

In general, misestimations can be divided into two categories. The noisy signal is either oversuppressed or undersuppressed. The costs of those two situations are sound class and goal dependent. The costs for quality and intelligibility improvement goals are discussed, though not quantified, for vowels, consonants, and silence segments below.

Consonant sounds are typically short-duration, and low-energy segments. They are

critically important in intelligibility tasks while they are not as important in quality improvement tasks. Oversuppression of the low-energy consonant sounds causes erasure of the consonant cues that are important for intelligibility. However, typically, oversuppression does not significantly degrade the perceptual quality.

Undersuppression leaves residual noise on the consonant sound that can create artificial cues in the signal and confuse the human auditory system. Therefore, undersuppression can also degrade intelligibility. Moreover, the residual noise can also degrade the perceptual quality.

Vowels and silence segments are the major determiners of the speech quality. Formant locations and bandwidths are the primary determiners of the quality of vowels. Oversuppression can change the locations of the formants and/or the formant bandwidth. Moreover, oversuppression can cause deletion of perceptually important harmonic frequencies as shown in Fig. 4. Both situations can substantially degrade the speech quality. Undersuppression can also change the formant locations and/or reduce the formant resolution in vowels. Moreover, undersuppression has an additional cost of leaving significant residual noise particularly at the nonharmonic, low-energy frequencies. Intelligibility of vowels are not considered in some intelligibility tasks such as DRT. However, formant transitions within a vowel can significantly impact the intelligibility of a neighboring consonant. Therefore, oversuppression and undersuppression of a vowel spectrum can degrade the intelligibility of the neighboring consonant.

Oversuppression has minimal cost for silence segments. In fact, oversubtraction is typically desirable, since there is no speech present in the signal. However, undersuppression typically has two costs. The first cost is the wideband residual noise and the second cost is the narrowband musical noise as shown in Fig. 4. Musical noise artifact in particular can significantly degrade the perceptual quality.

3.1.3 Quantifying the Probability of Misestimation

The $P(G_{i,j}/\hat{X}_{i,j}, \hat{G}_{i,j}, c, G_{i,j})$ term in Eq. 13 is hard to estimate in general because the statistics of speech and noise are not known at a time-frequency location in noisy speech.

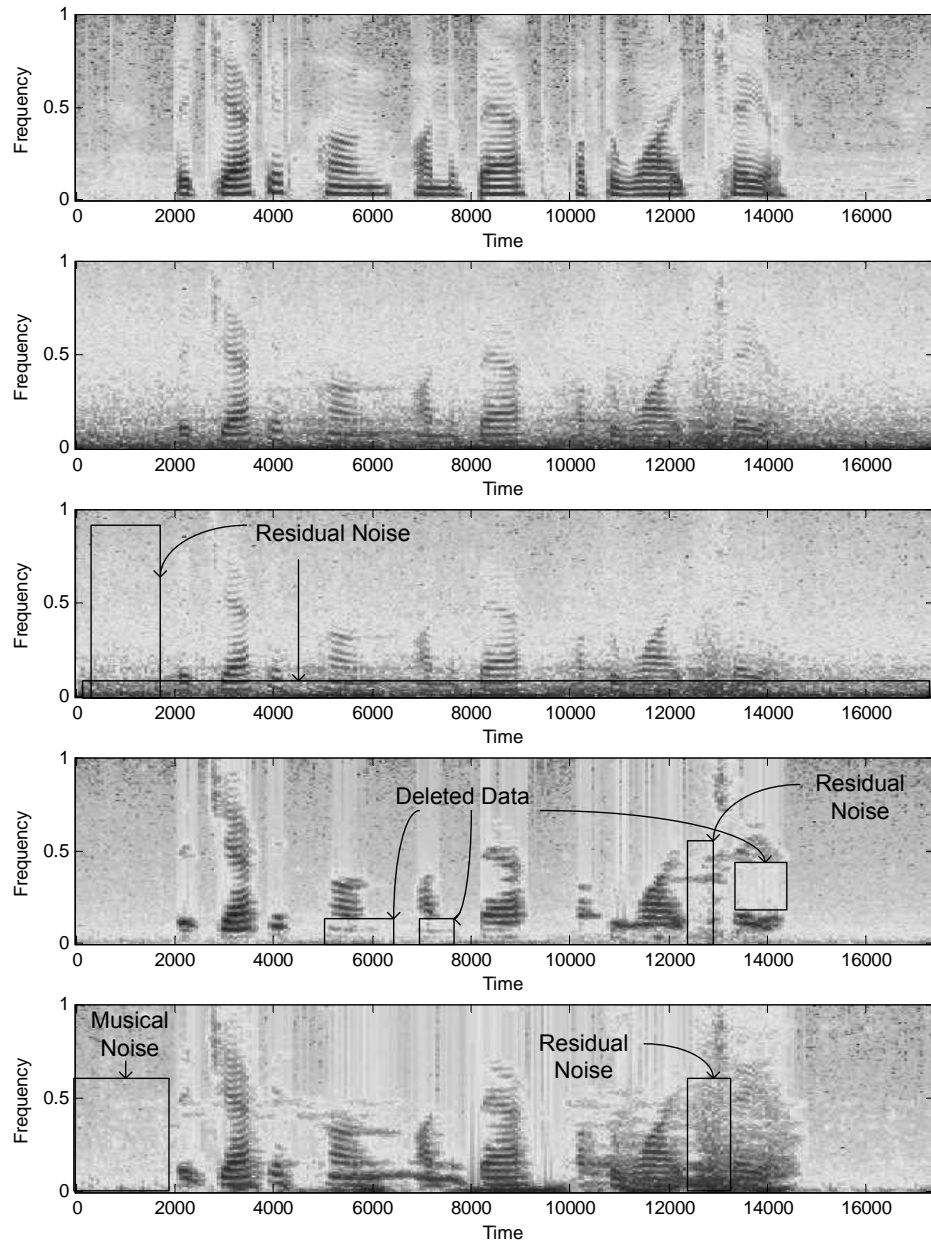


Figure 4: Effects of oversuppression and undersuppression is illustrated. Spectrogram of a clean speech utterance is shown in the first figure. The same signal is contaminated with 0 dB M2 tank noise in the second figure. The third figure shows the output of an MMSE estimator that attempts to clean the noisy speech. MMSE estimator with an oversuppression factor is used in the fourth figure. MMSE estimator with a less aggressive oversuppression factor is shown in the last figure.

Therefore, a heuristic measure is used to assess P_m as described below.

Two factors for misestimations, the uncertainty in the *a priori* parameters and the variance of the estimator, are identified in Section 3.1.1. In this section, these two risk factors are quantified for further analysis.

The Robust Bayesian Analysis (RBA) framework can be used to quantify the risks associated with the uncertainty in the prior parameter. There are three main RBA approaches to measure the robustness, or sensitivity, of the estimator to the uncertainty in the prior distribution. The first approach is called the informal approach where a few highly probable priors are considered, and the *a posteriori* means are compared to measure the robustness of the estimator. In the second approach, called the global robustness, range of *a posteriori* means of all possible prior functions is used to assess the robustness. Finally, the local robustness approach uses the derivative of the estimator with respect to the prior variable at the operating point to measure the rate of change of inferences, sensitivity, of the estimator to the prior.

Even if the *a priori* parameters were perfectly known, misestimations still occur because of the inherent uncertainty in the estimation process. This factor can be quantified for an estimator using the Monte Carlo simulation technique for given *a priori* parameters or analytical derivation when possible.

In this thesis, the local robustness approach used in RBA analysis is used to quantify both SNR and estimator uncertainty. The suppression gain in a typical speech enhancement estimator decreases with an increasing rate with decreasing SNR. Hence, the local derivative of the gain function with respect to SNR increases with decreasing SNR. Uncertainty of the estimator also increases with decreasing SNR. Therefore, the local derivative approach of the RBA framework can also be used as a measure to assess the uncertainty of the estimator at different SNRs.

3.2 Analysis of Spectral Subtraction Method Using the Proposed Framework

Bayesian estimators are commonly used for enhancing the noisy speech signal; however, some heuristically modified versions of the “optimal” estimators are shown to perform better in both ASR and perceptual quality enhancement tasks as discussed in Section 3.1. Performance improvement obtained with some of the modified systems can be analyzed under the proposed theoretical framework. Several variants of the popular spectral subtraction technique are analyzed below.

3.2.1 Analysis of the Modified Spectral Subtraction

In the spectral subtraction (SS) algorithm, clean speech power $|X|^2$ is estimated with

$$|\hat{X}|^2 = |Y|^2 - |\hat{N}|^2. \quad (14)$$

where $|\hat{X}|^2$ is the estimated power, $|Y|^2$ is the noisy signal power, and $|\hat{N}|^2$ is the estimated noise power. Therefore, the gain function of the optimum linear SS is

$$G_{ss} = \left(1 - \frac{1}{\Gamma}\right), \quad (15)$$

where Γ is *a priori* SNR and defined as

$$\Gamma = \frac{|Y|^2}{|\hat{N}|^2}. \quad (16)$$

Although the SS filter is an optimum ML filter, some of the heuristically modified versions of it are shown to generate higher quality speech compared to the SS system. Moreover, some SS-based nonlinear estimators, when used at the front-end of an ASR system, are shown to provide higher accuracy in noisy environments compared to the SS system. In the following two sections, popular examples to such modified SS systems and their superior performance compared to the optimum SS system are discussed using the proposed framework.

3.2.2 Spectral Subtraction for Perceptual Quality Enhancement

One of the variants of the linear SS algorithm is the modified SS (MSS) algorithm that uses an oversuppression factor to improve the perceptual quality [5]. The MSS system estimates

clean speech power with

$$|\hat{X}|^2 = |Y|^2 - \alpha|\hat{N}|^2 \quad (17)$$

where α is the oversuppression factor and $\alpha \geq 1$. The gain function of the MSS system is

$$G_{mss} = \max\left(1 - \frac{\alpha}{\Gamma}, \Theta\right) \quad (18)$$

where Θ is the minimum gain threshold. Typically, the transfer function G_{mss} is set to a constant, Θ , below an SNR threshold (Ex: 5.5 dB) as shown in Fig. 5. The α and Θ modifications to the SS algorithm are analyzed under the proposed framework in this section. The purpose of this analysis is to formally explain the better perceptual quality of the MSS system and set the theoretical groundwork for further improvements.

The local derivative method is used to measure the sensitivity of an estimator to the SNR misestimations as discussed in Section 3.1.3. The sensitivity function of SS (S_{ss}) with respect to SNR is

$$S_{ss} = \frac{\partial G_{ss}}{\partial \Gamma} = \frac{1}{\Gamma^2}. \quad (19)$$

The sensitivity function of MSS (S_{mss}) is

$$S_{mss} = \frac{\partial G_{mss}}{\partial \Gamma} = \begin{cases} \frac{\alpha}{\Gamma^2} & \text{if SNR} > \Theta, \\ 0 & \text{if SNR} \leq \Theta. \end{cases} \quad (20)$$

The gain functions and their sensitivity functions are shown in Fig. 5 for SS and MSS. Sensitivity of SS increases rapidly with decreasing SNR. Therefore, particularly at lower SNRs, the risk of musical noise, which have a high perceptual cost, is significantly higher compared to the higher SNRs. The MSS algorithm uses a constant suppression factor at those high risk SNRs to make the system more immune to sudden fluctuations. From an RBA stand-point, sensitivity of the MSS filter to the misestimations of SNR becomes zero since the MSS filter uses a constant suppression factor. Although P_m is higher compared to the SS filter, since SS is the optimum filter, enhanced speech contains less musical noise artifact and has higher perceptual quality in MSS.

At higher SNRs, P_m is relatively lower, but there is still a risk of leaving residual noise on the enhanced signal. The MSS system multiplies the estimate of the noise variance with

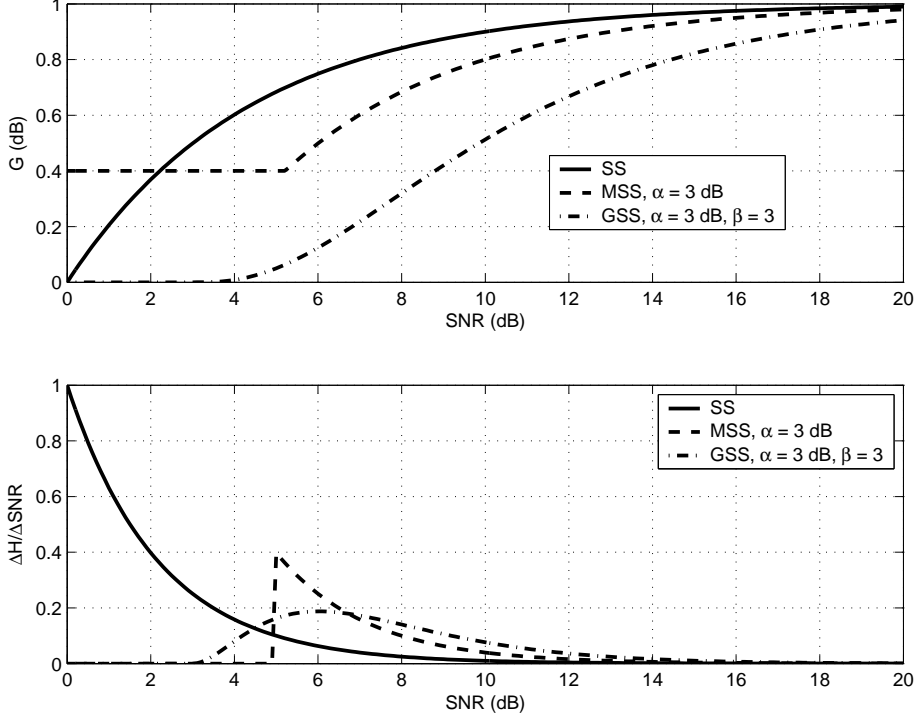


Figure 5: The gain functions H (dB) is shown in the top figure. Sensitivity of the gain functions to the prior parameters are shown in the bottom figure.

a constant α factor to achieve oversubtraction and reduce that risk. An important thing to note is that the amount of oversuppression increases with decreasing SNR which ensures more aggressive noise suppression for higher risk SNR regions.

The 5.5 dB threshold can be explained using the sensitivity function of the MSS system which increases slowly with decreasing SNRs until 6 dB starts to increase more rapidly below 5 dB as shown in Fig. 5. Therefore, 5.5 ± 0.5 dB is a critical threshold range for the sensitivity function.

One further modification to the SS system is the generalized spectral subtraction (GSS) system [27] that not only uses a constant α factor for oversubtraction but also uses a β exponent modification in the gain function. The β factor gives a second degree of freedom for adjusting the subtraction factor. The GSS gain is

$$G_{gss} = \left(1 - \frac{\alpha}{\Gamma}\right)^\beta. \quad (21)$$

The GSS gain and its SNR sensitivity is shown for $\alpha = 0$ dB and $\beta = 2$ in Fig. 5. The G_{gss} curve closely follows the G_{ss} curve at high SNRs while at lower, more sensitive

SNRs, it starts departing from the G_{ss} curve. Note that the SNR threshold is not needed in GSS, since, at low SNRs, the new system has considerably lower sensitivity compared to SS and MSS algorithms. As opposed to SS and MSS, sensitivity of GSS starts decreasing with decreasing SNR below around 5.5 dB which explains how it can avoid the musical noise artifact without any SNR threshold.

The higher performance of the GSS system compared to the MSS system is related to the fact that it has two degrees of freedom, and the suppression gain can be adjusted with a greater flexibility as explained above. Another dimension of flexibility can be obtained if the oversubtraction factor, α , is a function of the estimated SNR. The gain function with an SNR-dependent α is

$$G_{v\alpha} = \left(1 - \frac{\alpha(\Gamma)}{\Gamma}\right). \quad (22)$$

In [71], the spectrum is divided into four bands, and a linear function of SNR is used in each band to increase α for more aggressive suppression at lower SNRs. Although, MSS and GSS works with similar principals, the system in [71] provides the flexibility to tune α separately at each SNR which is shown to improve quality at low SNR environments.

3.2.3 Spectral Subtraction for Noise-Robust ASR

There are SS-based systems that are specifically designed for noise-robust ASR applications. The nonlinear SS (NSS) system proposed in [87] has a gain factor

$$G_{nss} = \left(1 - \frac{\Phi(\Gamma, \rho, |\hat{N}|^2)}{|\hat{Y}|^2}\right) \quad (23)$$

where the noise estimate Φ is

$$\Phi(\Gamma, \rho, |\hat{N}|^2) = \frac{\rho}{1 + \Gamma\rho}, \quad (24)$$

and ρ is a constant. Φ is not allowed to exceed $3|\hat{N}|^2$ or drop below $|\hat{N}|^2$, and the ρ factor is frequency dependent. The relatively good performance of the NSS filter in ASR tasks is discussed below using the proposed framework.

ASR systems commonly use Mel Frequency Cepstrum Coefficients (MFCCs) as discussed in Chapter 2. In the MFCC extraction algorithm, the first step is to calculate the logarithm

of the energy in each auditory subband. The masking effect and the winner-take-all effect of the nonlinear logarithm function can be used to explain the relatively good performance of the NSS system in ASR tasks. These two effects are discussed below.

The masking effect can be described as follows. For the additive and independent noise case, expected value of the log-energy of a noisy signal is

$$\log(Y^2) = \log((X + N)^2) \quad (25)$$

where X is the clean signal, N is the noise signal, and Y is the noisy signal. If $|N|^2$ is significantly greater than $|X|^2$, then the noise masks the signal, and

$$\log(Y^2) \approx \log(N^2). \quad (26)$$

The winner-take-all behavior of the logarithm function is a consequence of the masking effect and can be described as follows. Logarithm of the sum of N numbers can be approximated with a subset of those N numbers that are significantly higher than the rest of the numbers. Therefore, logarithm of the total energy in an auditory subband can be approximated using the total energy of the few high-energy frequencies that can mask the rest of the frequencies. Moreover, the high-energy frequencies typically have higher SNRs compared to the rest of the frequencies in an auditory channel. Hence, oversuppressing the lower SNR frequencies, where the estimator is less reliable and the speech energy is expected to be relatively lower, while operating close to the optimum SS system for high SNR frequencies that have higher speech energy enables relatively reliable extraction of the MFCC features in the NSS system.

There are spectral subtraction schemes, such as subband spectral subtraction (SSS) [102], that attempt to estimate the total signal power in a subband. It has been observed that the ASR performance improves when the total subband power is estimated as opposed to estimating the signal power at each frequency. In this case, the system has a better estimate of the SNR, and it performs better since the *a priori* parameter is more reliable. The uncertainty, hence the variability, of the SNR parameter decreases in this case because of the smoothing effect of using total power.

CHAPTER IV

AHARMONIC COMB FILTERING OF NOISY SPEECH

Comb filtering is commonly used in tandem with other speech enhancement algorithms for improving the perceptual quality [64]. The main idea is to leverage the harmonic structure of the voiced speech signal for suppressing the background noise. The noisy signal at the nonharmonic locations can be suppressed without distorting the perceptually important speech cues while suppressing significant amounts of noise [107], [24], [130].

There are several problems with the basic comb filtering technique. The first problem is that voiced speech is quasiperiodic even within a small analysis window. Therefore, the harmonicity assumption of the comb filter is not valid in general. Adaptive comb filters address this problem by finding the pitch epochs in a speech frame without a periodicity assumption [84], [47]. However, finding the correct pitch epochs is challenging at low SNRs and/or nonstationary noise environments.

Comb filtering can be done in the frequency domain as well as the time domain. In both cases, sampling resolution limits the effectiveness of the algorithm. The system proposed in [110] uses fractional sample delay to solve the problem in the time domain. However, the system in [110] is not adaptive, and it is susceptible to pitch epoch misestimations.

Frequency domain comb filters use windowing for analysis of a speech frame which can further reduce the resolution of the harmonics and create high energy side lobes in voiced speech spectrum. Combined with the quasiperiodicity and resolution problems, there might be more than one high signal power (HSP) locations around the harmonic frequency. Previous comb filtering techniques do not consider the high-energy sidelobes around the harmonic frequency which can significantly degrade the speech quality.

A sinusoidal model of the speech signal is proposed in [98] that enables aharmonic analysis of the voiced speech signal. However, accurate sinusoidal modelling of noisy speech is

challenging. Misestimations, for example, are found to degrade the speech quality significantly in sinusoidal coders [100] and speech enhancement systems using sinusoidal models [3].

HSP locations in the speech spectrum can be detected using the GEMS signal under all acoustic noise conditions and SNRs [7] as mentioned in Section 2.4. In this chapter, an aharmonic comb filtering (ICF) technique is proposed that uses the GEMS signal to detect HSP locations in the voiced speech spectrum without any harmonicity assumption. HSP location detection is robust to ambient noise because GEMS is a non-acoustic sensor. The ICF system is described below.

4.1 Aharmonic Comb Filter

4.1.1 The HSP Detection Problem

Speech production mechanism is commonly modeled with a linear filter driven by an excitation signal. Therefore, the speech signal

$$s(n) = h(n) * e(n) \quad (27)$$

is a convolution of the vocal tract filter $h(n)$ and the excitation signal $e(n)$. Excitation signal for voiced speech is typically modeled with a periodic pulse generator for voiced speech with a short-time spectrum

$$e_v(f_n) = \sum_{k=0}^{N-1} \delta(f_n - kF_p) \quad (28)$$

Hz where k is an integer number, F_p is the short-time average pitch frequency, and N is the analysis window length. $e_v(f_n)$ has harmonics at every kF_p Hz, when convolved with $h(n)$, create harmonics in the final speech spectrum as shown in Fig. 6.

The model described above, although frequently used in speech applications, is a rough approximation of the actual system. Harmonics in the voiced speech spectrum usually does not have equal space between the, and sidelobes with variable bandwidths are created around the harmonics as shown in Fig. 6.

The problems of the harmonic model of speech can significantly degrade the speech quality. Therefore, an aharmonic model of voiced speech spectrum is proposed that is

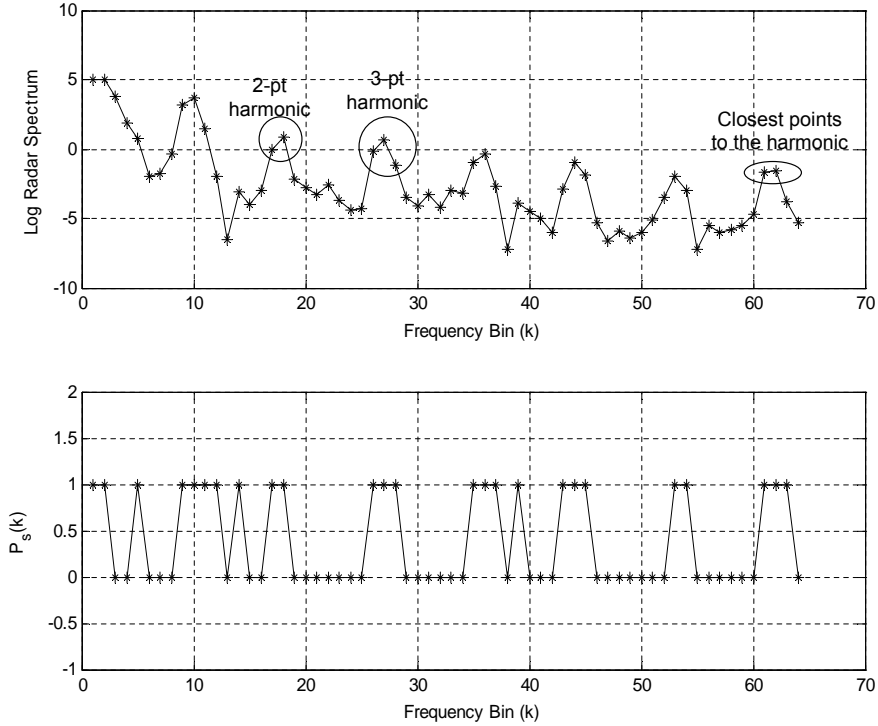


Figure 6: An example GEMS spectrum is shown in the top figure. The output of the HSP detector is shown in the bottom figure. The number of HSP points at a harmonic location is variable because of the bandwidth variability. In the GEMS spectrum HSP locations with two and three points are shown. Moreover, the resolution problem is clearly shown with a case where the exact harmonic position is in approximately in the middle of the two HSP locations.

defined by

$$S'_v(f_n) = \begin{cases} S_v(f_n) & \text{if } f_n \text{ is an HSP location,} \\ 0 & \text{if } f_n \text{ is not an HSP location.} \end{cases} \quad (29)$$

The aharmonic model does not make any periodicity assumption nor does it assume a strict harmonic structure for the voiced speech spectrum. Detection of HSP locations is discussed in the next section.

4.1.2 Detection of HSP Locations in Voiced Speech Spectrum Using the GEMS Device

The GEMS signal is correlated with the air flow velocity at the glottis during voiced speech [8]. Thus, if both speech and GEMS signals are windowed using the same window, then the GEMS spectrum can accurately indicate the (HSP) locations in the voiced speech spectrum.

Moreover, GEMS-based HSP location estimation is independent of the ambient noise since the GEMS signal is immune to acoustic noise.

The GEMS spectrum is divided into subbands with a bandwidth of pitch frequency, and HSP locations are detected independently in each band. Pitch information is extracted from the GEMS signal. An autocorrelation method is used at this step where the maximum-correlation lag in the interval of 2.5 msec and 10 msec is chosen as the pitch.

A hard-decision thresholding algorithm is used for detecting the HSP locations, and the binary decisions are stored in a masking vector P_s . Two types of HSP locations are identified in a given subband:

1. A harmonic location typically covers at least two frequency samples because of the resolution problem. Therefore, $P_s(k)$ is set to 1 at the two highest energy frequencies in the subband to detect. The two closest neighbors of the harmonic frequency.
2. An harmonic location can be composed of more than two points as shown in Fig. 6. Therefore, a derivative rule is used to detect locations that are neighboring an HSP location and have significantly higher energy compared to a neighboring non-HSP location. The energy comparison is done as follows. $P_s(k)$ is set to 1 if

$$(\zeta_k - \zeta_{k-1}) - (\zeta_{k+1} - \zeta_k) > \zeta_{th} \quad (30)$$

$(k + 1)$ is a type-1 HSP location, and $(k - 1)$ is a non-HSP location. ζ_k denotes the natural logarithm of the signal energy at frequency k . ζ_{th} is an energy threshold and discussed below. $P_s(k)$ is also set 1 if

$$(\zeta_k - \zeta_{k+1}) - (\zeta_{k-1} - \zeta_k) > \zeta_{th} \quad (31)$$

$(k - 1)$ is a type-1 HSP location, and $(k + 1)$ is a non-HSP location. ζ_k denotes the natural logarithm of the signal energy at frequency k . The algorithm for setting ζ_{th} is discussed below.

For each subband in a windowed speech frame, a two iteration procedure is followed. In the first iteration, type 1 locations are found, and in the second iteration, type 2 locations are found.

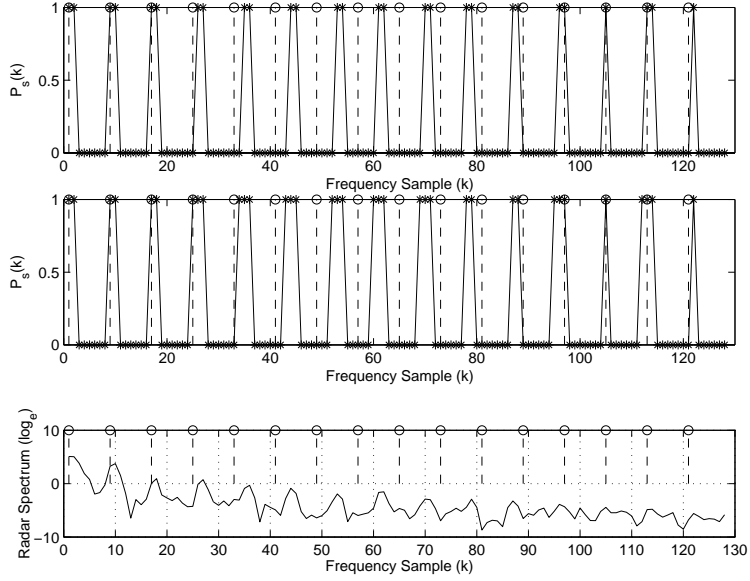


Figure 7: The GEMS spectrum is shown in the bottom figure. It is divided into subbands with a bandwidth of pitch. At each subband, the two iterations of the HSP detector is applied. The HSP locations detected in the first iteration is shown in the top figure. The HSP locations detected after the two iterations are shown in the middle figure.

ζ_{th} is initialized to 3 dB for each subband and adjusted adaptively. The HSP detection algorithm attempts to find at least three HSP locations in each subband. If the number of HSP locations is less than three for some bands after the two iterations, then ζ_{th} is decreased with a step size of 0.5 dB, and the second iteration is repeated for those bands until at least three HSP locations are detected or ζ_{th} is less than 0.

The initial value of ζ_{th} and the step size are found by observing the range of ζ which is typically between -10 and $+10$ for voiced GEMS signals. Below 3 dB, some of the HSP locations could not be detected which degraded the quality for some cases. The quality does not change when ζ_{th} is above 3 dB. The step size of 0.5 dB is set to minimize the computational complexity. Above 0.5 dB, some harmonics cannot be detected, and quality degradation is observed for some cases.

An illustration of the algorithm is shown in Fig. 7. The HSP locations detected by the first iteration are shown in Fig. 7-a. The HSP locations detected by the second and first iterations are shown in Fig. 7-b. Elements of the masking vector P_s that are not explicitly set to 1 by the algorithm are by default 0.

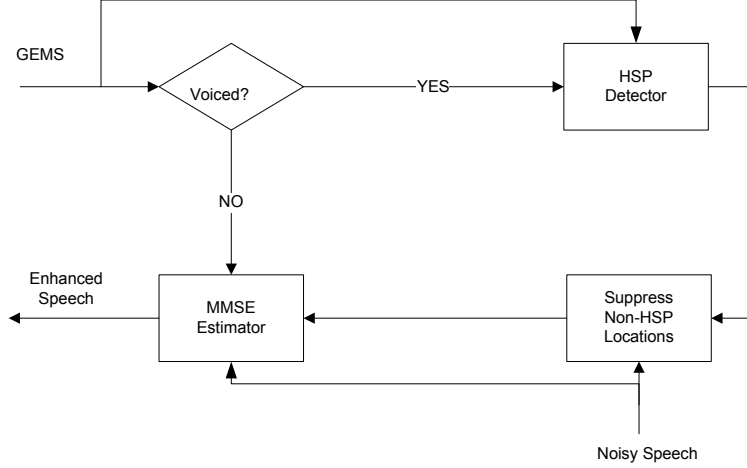


Figure 8: System overview of the aharmonic comb filter.

4.2 Speech Enhancement Using the Proposed HSP Detector

The system overview is shown in Fig. 8. The ICF system is used in tandem with an MMSE estimator [40] for voiced speech while only the MMSE estimator is used for unvoiced speech. HSP locations are estimated at the ICF block described in the next section. Non-HSP locations are severely suppressed with a constant suppression factor ($G_{min} = 10^{-3}$) while the HSP locations are enhanced with the MMSE filter. The suppression factor of the MMSE filter at frequency k is

$$G_k = \frac{\zeta_k}{1 + \zeta_k} \exp\left(\frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right) \quad (32)$$

where ζ_k is the SNR,

$$v_k = \frac{\zeta_k}{1 + \zeta_k} \lambda_k, \quad (33)$$

and

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_y(k)} + \frac{1}{\lambda_n(k)} \quad (34)$$

where $\lambda_y(k)$ is the noisy signal power, and $\lambda_n(k)$ is the noise power.

The autocorrelation based periodicity feature described in Section 2.4.2.1 is used as the measure of voicing in a speech frame. Since voicing detection is done using the GEMS signal, voicing detection is accurate at all noise conditions. Hard thresholding is used for voicing detection. If voicing is above a hard-threshold, the speech frame is classified as voiced, and if voicing is below the hard-threshold, the speech frame is classified as unvoiced.

4.3 Experiments

An A/B subjective quality test and two objective measure tests are used to compare the performance of the algorithm with the MMSE estimator. Harvard sentences from the ARCON database are used for the A/B test. Four male and four female speakers are used with the Blackhawk helicopter noise and the M2 tank noise under severe noise conditions. A total of 16 utterances are used in the test. Seven male and seven female listeners took the test. Among the listeners, five males and five females are native English speakers. Subjects chose A, B, or same options based on the quality of the enhanced speech utterance and described the reason for their choice. The question that was directed to the subjects was “Which sample sounds more pleasant?”.

The A/B subjective quality test results are shown in Table 2. Sixty percent of the time the proposed system was chosen as the better quality system. Thirteen percent of the time the MMSE system was favored, and twentyseven percent of the time listeners thought the quality is the same. Confidence intervals of the subjective test results are shown in Table 3. Student’s t-test is used to calculate the confidence intervals with a 95 percent confidence level. The subjective test results show that suppressing the low-power frequencies in the voiced speech spectrum while preserving the perceptually important HSP locations using the GEMS device significantly improves the speech quality.

Log-likelihood ratio and segmental SNR measures, which are commonly used objective measures, are used to objectively assess the performance of the proposed system. The segmental SNR (S-SNR) measure and the log-likelihood ratio measure, using the LP coefficients, are defined in Chapter 2. The M2 tank noise and the Blackhawk helicopter noise were used for testing. Fifteen minutes of speech from the DRT test set of ARCON database was hand-labeled and segmented into five voiced phonetic classes and one unvoiced/non-speech sound class for the objective measure tests.

The voicing threshold of 0.7 proposed in [44] is found to perform well in the experiments. The system is also tested with threshold values above and below 0.7. Speech quality did not degrade noticeably for the tested utterances if the threshold values are between 0.6 and 0.8. Below 0.6, musical noise is observed because unvoiced speech is enhanced using the

GEMS signal which does not contain information for unvoiced speech. Above 0.8, sound quality starts degrading because some of the voiced speech frames are not enhanced with the proposed system.

4.3.1 Results

Table 2: Results of the A/B speech quality tests.

| Preference | Percentage |
|-----------------|------------|
| Proposed System | 60 |
| MMSE system | 13 |
| Same | 27 |

Table 3: 95 percent confidence intervals of the A/B test results using student's t-test.

| Preference | Confidence Interval |
|-----------------|---------------------|
| Proposed System | [56, 64] |
| MMSE system | [8, 17] |
| Same | [21,32] |

The segmental SNR and the log-likelihood ratio results for the M2 tank noise are shown in Tables 4 and 5 respectively. Performance is compared for the voiced classes only since the ICF system is used for voiced sounds. The proposed system has higher S-SNR for all sound classes and both noise types. Significant difference in the log-likelihood ratio measure is not observed in the proposed system compared to the baseline system. Similar results were obtained for the Blackhawk helicopter noise as shown in Tables 6 and 7.

Table 4: Comparison of the proposed system with the noisy system using the segmental SNR measure for the M2 tank noise. Highest SNR cases are shown with bold fonts.

| Phone Class | Noisy (dB) | Baseline (dB) | Proposed System (dB) |
|------------------|------------|---------------|----------------------|
| Voiced Fricative | -8.9 | -5.0 | -4.6 |
| Voiced Plosive | -8.8 | -4.3 | -3.5 |
| Vowel | -1.5 | 2.6 | 3.2 |
| Semivowel | -5.5 | -0.6 | 0.2 |
| Nasal | -7.8 | -1.9 | -1.0 |

Table 5: Comparison of the proposed system with the noisy system using the log-likelihood ratio measure for the M2 tank noise.

| Phone Class | Noisy | Baseline | Proposed System |
|------------------|-------|----------|-----------------|
| Voiced Fricative | 1.3 | 0.9 | 0.9 |
| Voiced Plosive | 1.6 | 1.1 | 1.1 |
| Vowel | 1.0 | 0.6 | 0.5 |
| Semivowel | 2.2 | 1.5 | 1.4 |
| Nasal | 1.5 | 0.9 | 0.9 |

Table 6: Comparison of the proposed system with the noisy system using the segmental SNR measure for the Blackhawk helicopter noise. Highest SNR cases are shown with bold fonts.

| Phone Class | Noisy (dB) | Baseline (dB) | Proposed System (dB) |
|------------------|------------|---------------|----------------------|
| Voiced Fricative | -9.4 | -5.6 | -4.8 |
| Voiced Plosive | -9.0 | -5.3 | -3.8 |
| Vowel | -3.0 | 2.8 | 3.0 |
| Semivowel | -6.3 | -0.3 | 0.1 |
| Nasal | -8.4 | -2.4 | -1.9 |

Table 7: Comparison of the proposed system with the noisy system using the log-likelihood ratio measure for the Blackhawk helicopter noise.

| Phone Class | Noisy | Baseline | Proposed System |
|------------------|-------|----------|-----------------|
| Voiced Fricative | 1.61 | 1.23 | 1.29 |
| Voiced Plosive | 2.18 | 1.90 | 1.90 |
| Vowel | 1.15 | 0.99 | 0.99 |
| Semivowel | 2.50 | 2.35 | 2.35 |
| Nasal | 1.96 | 1.69 | 1.69 |

CHAPTER V

SPARSE REPRESENTATION OF VOICED SPEECH SPECTRUM FOR NOISE-ROBUST LPC EXTRACTION

Linear prediction (LP) is a widely used parametric model of the speech spectrum that creates a perceptually attractive model of the spectral envelope partly because it models the perceptually important spectral peaks more accurately than the spectral valleys [128]. However, in additive noise environments, resolution of the spectral peaks in the linear prediction (LP) spectrum decrease which can significantly reduce the quality of the coded speech [73]. Moreover, performance of the LP-based ASR systems drop substantially with increased background noise [114]. Therefore, it is important to develop noise-robust linear prediction coefficients (LPC) extraction algorithms that are immune to background noise. Such algorithms can find applications in all speech applications, such as parametric speech coders [128] and automatic speech recognition (ASR) [114], where the LP method is used.

There are noise-robust LPC extraction systems that perform relatively well for some noise types and Signal-to-Noise-Ratios (SNR). The system proposed in [72] uses iterative MAP estimation to estimate the LPC parameters in noisy environments. EM algorithm is used for ML estimation of the LPC parameters in [33]. A subspace method is proposed in [22] that generates mean pole locations that are close to correct locations while having small error variance.

In the previous chapter, an aharmonic comb filter (ICF) is proposed to enhance noisy speech. In this chapter, noise-robust extraction of LPC parameters using the ICF system is discussed. In a noisy environment, suppressing the signal at the non-HSP frequencies can increase the accuracy of the LPC parameters because the LP method can still model the spectral envelope of voiced speech when a significant portion of the spectrum is severely suppressed. For example, if pitch and LPC order are low enough, LPC can generate a smooth spectral envelope of voiced speech using the envelope information available only at

the harmonic frequencies. A more detailed analysis of the effect of sparse spectrum on LPC extraction is done below. Noise-robust LPC extraction using the aharmonic comb filter is discussed in Section 5.1.1. Experimental results are presented in Section 5.2.

5.1 *Effect of Sparse Spectrum on LPC Extraction*

Using the LPC parameters $a_{j,k}$ at speech frame k , the speech signal

$$s_k(n) = \sum_{j=1}^L s_k(n-j)a_{j,k} + e_k(n) \quad (35)$$

where L is the LPC order, and $e_k(n)$ is the residual signal. The LPC parameters, $a_{j,k}$, can be derived by minimizing the total energy of the residual noise

$$\epsilon_k = \sum_{n=-\infty}^{\infty} (e_k(n))^2 \quad (36)$$

Using Eq. 35, the speech spectrum

$$S_k(z) = E(z)H_k(z) \quad (37)$$

where

$$H_k(z) = \frac{1}{1 - \sum_{k=1}^L a_{j,k}z^{-k}}. \quad (38)$$

Using Eq. 37 and the Parseval theorem, the spectral domain equivalent of Eq. 36 is

$$\epsilon_k = \int_{-\pi}^{\pi} \frac{|S_k(z)|^2}{|H_k(z)|^2}. \quad (39)$$

An interesting property of the LP method is that it can generate a smooth spectral envelope of a sparsely represented spectrum such as in the case of voiced speech [128]. One way of understanding the way LPC handles the sparse spectrum is by analyzing the error function in Eq. 39. In a sparse spectrum, some of the frequencies are severely suppressed and $S_k(z)$ is significantly smaller than the spectral envelope $H_k(z)$. In this case, the ratio in the error term at those frequencies are significantly smaller compared to the frequencies that are not suppressed. Therefore, the LPC extraction algorithm becomes insensitive to those suppressed parts of the spectrum, and LPC fits a model based on the available spectrum. As long as the LPC order is low enough and the sampling rate of the spectrum is high

enough, LPC does not follow the fine structure of the spectrum. Hence, the resulting LPC spectrum is usually a good approximation of the LPC spectrum with the complete data for the speech signal.

In the next section, use of sparse spectral representation for noise-robust extraction of the LPC parameters from voiced speech is described.

5.1.1 Noise-robust LPC Extraction Using the Aharmonic Comb Filter

Voiced speech signal is typically modeled with a smooth all pole transfer function $h(n)$ driven by a quasistationary source signal $e(n)$. The spectrum of $e(n)$ is similar to an impulse train where each impulse occurs at the integer multiple of the fundamental frequency F_0 . Thus, the voiced speech spectrum $S(k)$ is a sampled version of $H(k)$ at the HSP locations. Although there is still some energy at the non-HSP locations, it is negligible compared to the energy at the HSP locations for clean speech as shown in Fig. 10.

While the lack of spectral information at the non-HSP locations typically does not affect the operation of LPC, the effect of relatively low power at those frequencies should be taken into account in noisy environments. Non-HSP locations typically have significantly lower SNRs compared to HSP locations in a noisy environment which can significantly distort the LPC spectrum. However, this problem can be ameliorated if the HSP locations are detected and the signal at the non-HSP locations are severely suppressed. The ability of LPC to handle such sparse spectral representations can improve the LPC spectrum in noisy environments especially if noise is concentrated in the low frequency bands where most of the speech signal resides.

The ICF system described in the previous chapter is used to detect and suppress the non-HSP locations in voiced speech as shown in Fig. 9. The speech signal spectrum $S(k)$ is labeled voiced or unvoiced using the voicing detector described in Section 4.2. If the frame is labeled as voiced, then the ICF system suppresses the non-HSP locations, and LPC parameters are extracted from the sparse spectral representation.

Fig. 10 shows a case where the spectrum of the voiced speech is contaminated with the Blackhawk helicopter noise at 0 dB, and the LPC spectrum of the noisy speech is highly

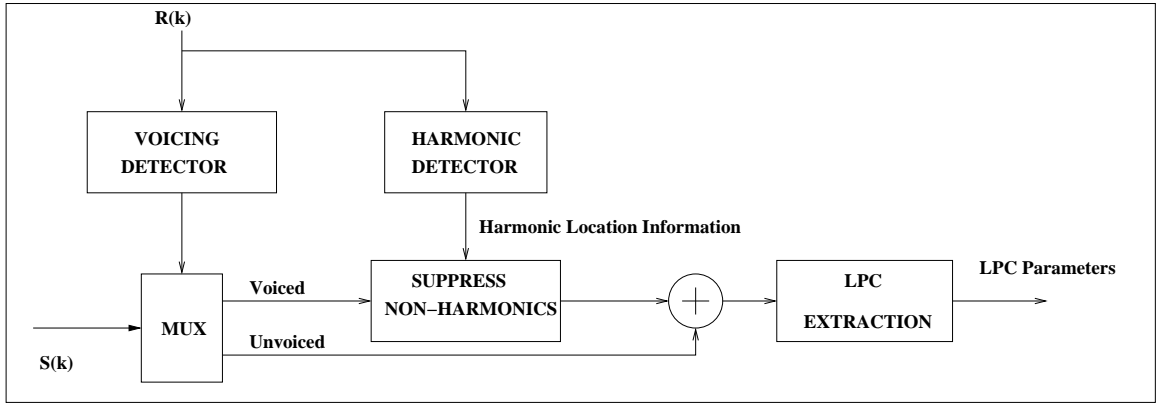


Figure 9: Overview of the proposed LPC extraction system. $S(k)$ is the spectrum of the speech signal, and $R(k)$ is the spectrum of the radar signal.

distorted compared to the clean spectrum. Resolution of the formant frequencies in the LPC spectrum significantly degrades for the noisy signal in Fig. 10 where all three peaks at the formant locations are smoothed because of noise. When non-HSP locations are detected and suppressed, a sparse spectral representation of the noisy speech is obtained. Fig.10-b illustrates the resulting spectrum for the case shown in Fig. 10-a. The LPC spectrum of the sparse spectrum resembles the clean speech spectrum in the first formant region. Moreover, the third formant is more prominent relative to noisy speech although its position is shifted compared to the clean case. Furthermore, the overall spectral level is much closer to the clean case when the non-HSP locations are suppressed. Thus, suppressing the non-HSP locations without any further spectral enhancement significantly improves the LPC spectrum in this example.

5.2 Experiments

Fifteen minutes of 8 kHz speech samples from the DRT sequences in the ARCON database have been hand-labeled into five phonetic classes that represent voiced phonemes. Itakura-Saito (IS) objective measure is used to compare the performance of the proposed system with the noisy system for each class. Background noise is added electronically to clean speech. Segmental SNRs for both environments are 0 dB in these experiments.

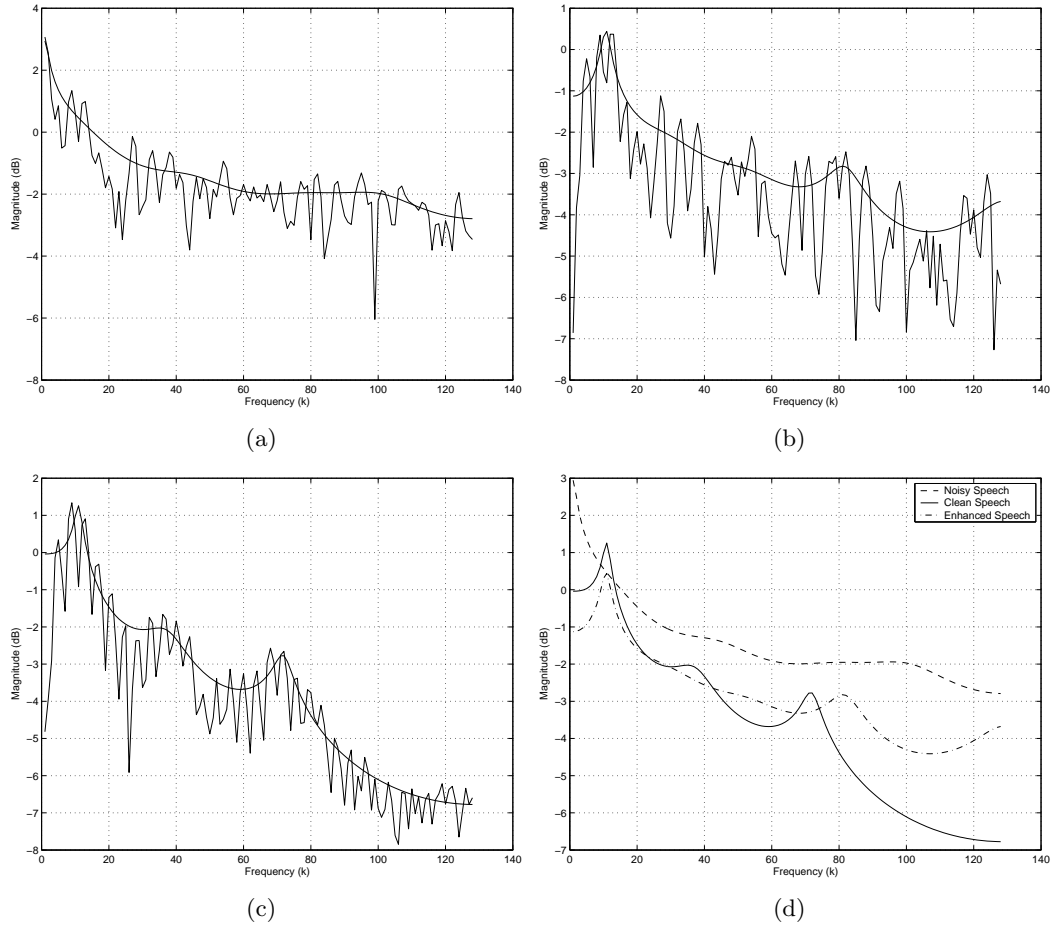


Figure 10: Illustration of noise-robust LPC extraction with the proposed system. The noisy spectrum and its LPC spectrum are shown in (a). The noisy spectrum processed with the proposed system and its LPC spectrum are shown in (b). The clean spectrum and its LPC spectrum are shown in (c). Comparison of the LPC spectra of the clean spectrum, noisy spectrum, and the processed spectrum is shown in (d).

Table 8: Performance results in terms of Itakura-Saito distortion measure. LPC spectrum is created with the full-spectrum in the baseline system. Sparse spectrum is used for LPC extraction in the proposed system. M2 tank noise used.

| Phone Class | Baseline | Proposed System |
|------------------|----------|-----------------|
| Voiced Fricative | 5.4 | 2.8 |
| Voiced Plosive | 5.3 | 3.2 |
| Vowel | 5.7 | 4.3 |
| Semivowel | 5.9 | 3.4 |
| Nasal | 5.4 | 2.9 |

Table 9: Performance results in terms of Itakura-Saito distortion measure. LPC spectrum is created with the full-spectrum in the baseline system. Sparse spectrum is used for LPC extraction in the proposed system. Blackhawk helicopter noise is used.

| Phone Class | Baseline | Proposed System |
|------------------|----------|-----------------|
| Voiced Fricative | 6.4 | 3.4 |
| Voiced Plosive | 6.5 | 3.5 |
| Vowel | 6.8 | 4.7 |
| Semivowel | 7.0 | 3.9 |
| Nasal | 6.6 | 3.2 |

The results are shown for the M2 tank noise and the Blackhawk helicopter noise environments in Tables 8 and 9 respectively. The proposed system reduces the mean IS distance significantly for all phoneme classes and both noise types. An interesting point to note is the relatively smaller gain for the vowel class. Vowels have relatively higher energy, and they are more immune to additive noise compared to other voiced sounds.

5.3 Conclusion

The ICF system described in the previous chapter is used for noise-robust extraction of the LPC parameters. A sparse spectrum is created by suppressing the non-HSP frequencies using the ICF system. The LPC extraction algorithm is theoretically shown to be robust to such sparse spectral information. The objective measure test using the Itakura-Saito spectral distance shows significant improvement in the LPC parameters compared to the noisy case. The proposed system can be used to increase the noise immunity of all speech applications that use the LPC parameters such as ASR and speech coding.

CHAPTER VI

SEGMENTATION-BASED ENHANCEMENT OF THE NOISY SPEECH SIGNAL FOR INTELLIGIBILITY IMPROVEMENT IN MELP SPEECH CODERS

Despite significant progress in improving the robustness of speech coders in noisy environments, intelligibility of parametric speech coders still decreases significantly in harsh noise environments [117]. The problem is particularly important for military applications where clearly understanding the correct message is critical. In this work, the intelligibility problem of a 2400 bps MELP coder in harsh noise environments is addressed using a front-end noise suppression system.

Noise-robust signal processing systems for parametric speech coders can be grouped into three broad categories. The first category of methods uses acoustic properties, such as microphone design including microphone arrays. In harsh-noise environments, or in situations where such elaborate measures are not practical, other processing must be performed. The second category of systems attempts to estimate clean speech parameters from the noisy speech signal [74],[148],[93],[128],[59],[22],[92]. The third category of systems attempt to suppress the noise in the speech signal before extracting the speech parameters. For example, the state-of-the-art 2400 bps MELPe speech coder uses this approach [7], as do systems proposed in [1], [141], [19], [55], and [91].

Among the noise suppression techniques, there are methods that leverage auxiliary sensors in addition to a noisy acoustic microphone channel[65],[28],[117]. In [65], a glottal correlation filter was designed using an auxiliary device (GEMS), and significant intelligibility improvement over the MELPe coder as measured by the Diagnostic Rhyme Test (DRT) was reported. However, most of the improvement was obtained by enhancing transmission of the voicing feature. The system proposed in [117] fused a GEMS signal, a throat-mounted,

gel-suspended accelerometer (p-mic) signal, and a bone-coupled accelerometer signal with an acoustic microphone signal. The fusion algorithm used the estimated quality of each signal in various frequency bands. The system described in [137] fused a high-pass filtered accelerometer signal mounted on the throat with a low-pass filtered noise cancelling microphone signal to improve intelligibility of coded speech in harsh-noise environments. This same system was extended in [138] by using additional sensors with different frequency band configurations.

Most of the previous statistical speech enhancement systems lack the ability to exploit the acoustic-phonetic properties of speech sounds. There have been efforts to include such *a priori* knowledge of the speech signal into the enhancement algorithm. In [42], a Hidden Markov Model (HMM) was trained using clean speech, and the noisy speech was segmented into stationary states using ergodic HMM. Although those systems were found to perform improvement in quality, they still do not fully exploit the acoustic-phonetic knowledge of speech since the HMM states do not necessarily correspond to any phonetic events. The system proposed in [34] had five states that represented five phoneme classes. The speech signal is segmented into one of those phoneme classes, and a class specific filter is applied to each segment. The system has inaccurate segmentation problems, does not make any distinction between voiced and unvoiced speech, and it is not designed for an intelligibility improvement task.

A Bayesian risk minimization (BRM) framework is proposed for speech enhancement in Chapter 3. One of the key aspects of the BRM framework is that each sound class should be enhanced based on its acoustic-phonetic properties and the goal of the enhancement system. In this chapter, a speech enhancement system is proposed that uses the acoustic-phonetic properties of speech. Speech is segmented into a number of broad-level sound classes, and each sound-class (e.g. vowels) is treated differently based on its acoustic-phonetic properties.

The signal presence detection framework [15],[75], [96] is used to detect perceptually important speech cues. These important regions of the spectrum are mildly suppressed while the rest of the spectrum is more aggressively suppressed to minimize the risk of deleting low-energy speech cues that are vital for intelligibility.

Previously reported systems that use the signal presence detection idea typically use an SNR -based signal presence measure that becomes less reliable with decreasing SNR . They are mostly targeted at increasing the subjective quality of the speech signal, which is primarily related to vowel and silence segments. Oversuppressing the low- SNR frequency regions typically emphasizes the formant frequencies in vowels and reduces the residual and tonal noise. However, oversuppression runs the risk of erasing the acoustic cues for low-energy consonant sounds that can degrade the intelligibility as discussed in Chapter 3. The proposed system reduces this risk by exploiting the acoustic-phonetic knowledge of the speech and the auxiliary sensor signals in addition to the estimated SNR .

This chapter is organized as follows. The challenge of SNR -based signal presence detection is discussed in Section 6.1. Auxiliary sensors that are used in this work are briefly described in Section 6.2. The proposed system is described in Section 6.3. The general methods that are used for signal presence detection are explained in Section 6.4, and the combination of methods that are used for each sound class are described in Section 6.5. Experimental results are presented and discussed in Section 6.6. The results are summarized in Section 6.7.

6.1 Problem Definition

In frame-based, linear speech enhancement systems, the clean speech signal at frame k and frequency i is estimated by

$$\hat{X}(k, i) = Y(k, i)G(k, i) \quad (40)$$

where \hat{X} is the estimate of the clean speech signal X , Y is the noisy speech signal, and G is the suppression factor. G is typically estimated using estimated SNR which is defined as

$$\zeta(k, i) = 10 \log_{10} \left(\frac{|\hat{X}(k, i)|^2}{\hat{\sigma}_{N(k, i)}^2} \right) \quad (41)$$

where $\hat{\sigma}_N^2$ is the estimated noise variance.

The speech signal does not exist at all points in the time-frequency plane in a noisy speech utterance. Therefore, probability of signal presence (P_s) can be estimated and used

to modify the gain function. Typically,

$$P_s = f(\zeta), \quad (42)$$

and the new gain function is

$$G'(k, i) = f(G, P_s) \quad (43)$$

where $f(\cdot)$ and $g(\cdot)$ are nonlinear functions. The general goal is to decrease G to oversuppress the noisy signal when P_s gets lower.

The signal presence detection can measurably improve speech quality. However, intelligibility of speech is dependent on low-energy consonant cues that can be significantly distorted with SNR -based oversuppression. The low-energy cues typically have relatively lower SNR in a noisy environment, and SNR -based oversuppression has a high risk of erasing these important speech cues as described in Chapter 3.

The problem of intelligibility is mainly related to the inadequacy of the SNR information in detecting the presence of low-energy, perceptually important consonant cues as discussed above. Two methods to detect signal presence in addition to the SNR -based detection are described below.

6.2 *Auxiliary Sensors*

Two auxiliary sensors, the gel-suspended accelerometer (p-mic) [117] and the general electromagnetic sensor (GEMS) device [8], are used in addition to the acoustic microphone in this chapter.

The GEMS device is described in Chapter 2. The p-mic device is a piezo-electric-based, gel-suspended vibrometer [124]. It can detect the vibrations on the skin during voiced speech if placed on the throat. The p-mic device is used in the segmentation block of the proposed system.

Both the GEMS and the p-mic devices provide voicing information. However, the GEMS spectrum typically has a higher bandwidth than the p-mic spectrum as shown in Fig. 11. Although the GEMS signal is robust to ambient noise, it can be noisy in cases of excessive skin fat and/or sensor misplacements. Therefore, the p-mic signal is used to complement

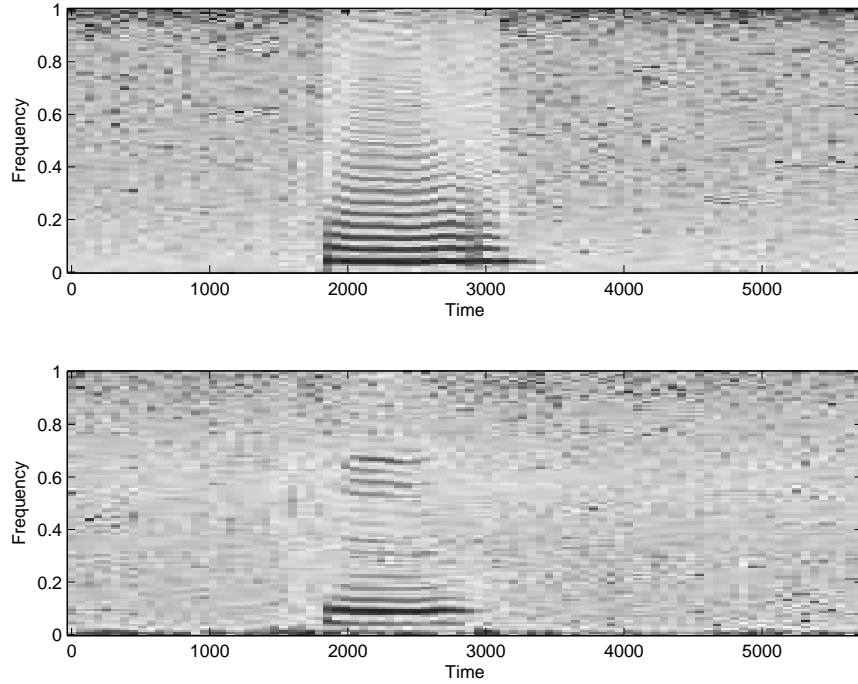


Figure 11: A spectrogram of the GEMS signal is shown in the top figure. A spectrogram of the p-mic signal is shown in the same figure for the same syllable sound. Sampling rate is 8 kHz for both signals.

the GEMS signal in the segmentation algorithm.

6.3 Proposed System

The general overview of the proposed system is shown in Fig. 12. The segmentation block uses the noisy speech signal, the GEMS signal, and the p-mic signal to segment noisy speech into broad-level acoustic-phonetic classes. The signal presence detection block uses the segmentation information, the high-signal-power (HSP) information for voiced speech, and the SNR estimates to evaluate the signal presence probability P_s . The MMSE estimator uses P_s to achieve oversuppression at the frequencies where speech has a low probability of existence.

The SNR estimation block uses spectral subtraction to estimate the clean speech spectrum. The noise spectrum is estimated by assuming stationary noise and averaging the spectra of the first ten noise-only frames of a noisy speech utterance using the segmentation information. Given the estimate of the noise power spectrum at frequency k and time frame

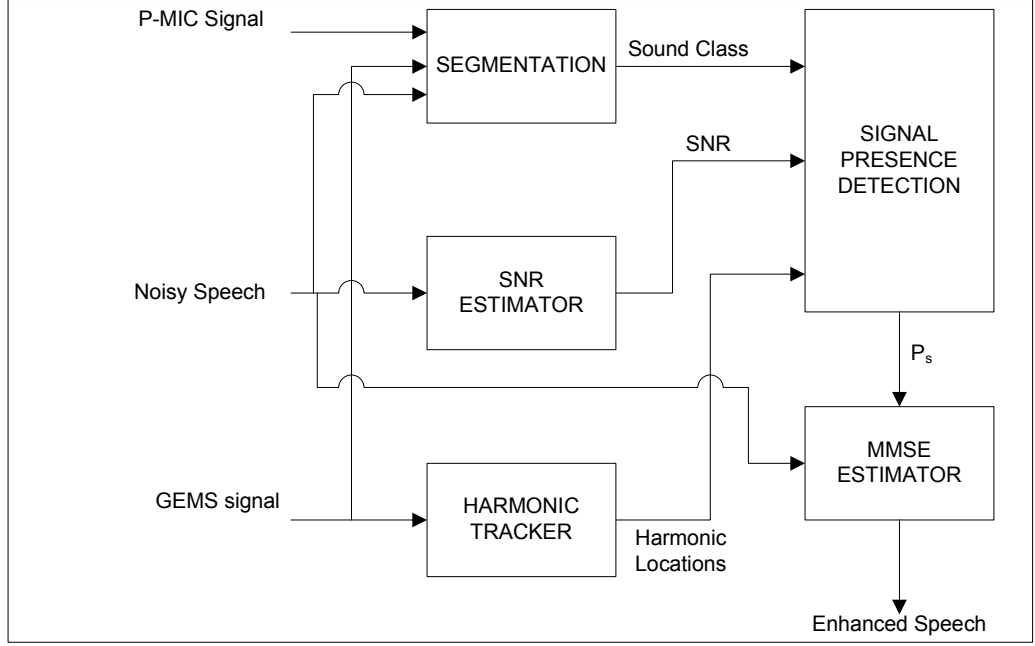


Figure 12: An overview of the proposed system. The speech signal is segmented using the noisy acoustic signal, the GEMS signal, and the p-mic signal. The segmentation information is used in addition to the SNR information and HSP locations for voiced speech to detect signal presence. Signal presence probability (P_s) is used at the speech enhancement block.

i , $\hat{\sigma}_N^2(k, i)$, and using the spectral subtraction technique, the estimate of the signal power is

$$|\hat{X}(k, i)|^2 = |Y(k, i)|^2 - \hat{\sigma}_N^2(k, i) \quad (44)$$

where $|Y(k)|^2$ is the noisy signal power. The SNR estimate is

$$\hat{\zeta}(k, i) = 10 \log_{10} \left(\frac{|\hat{X}(k, i)|^2}{\hat{\sigma}_N^2(k, i)} \right) \quad (45)$$

The HSP detector is discussed in detail in Chapter 4. The segmentation algorithm and the MMSE estimators are discussed below.

6.3.1 The Segmentation Algorithm

The segmentation algorithm detects the sound class of a speech segment based on hard-decision thresholding of the incoming sensor signals and noisy speech [30]. The noisy speech is segmented into vocalics, unvoiced fricatives, unvoiced plosives, transitional sounds, transients, voiced plosives, voiced fricatives, and silence. Transitional sounds include liquids, glides, nasals, and consonant to vowel transitions. Vocalics are highly voiced and relatively

Table 10: Combinations of the acoustic-phonetic, HSP location, and *SNR* information that are used for signal presence detection are illustrated for each sound class.

| Sound Class | Acoustic-phonetic | HSP | <i>SNR</i> | None |
|---------------------|-------------------|-------------------|------------------|------------------------------|
| Vocalics | X | 0 – 4 <i>kHz</i> | X | X |
| Unvoiced Fricatives | 0 – 2 <i>kHz</i> | X | X | 2 <i>kHz</i> – 4 <i>kHz</i> |
| Unvoiced Plosives | 0 – 400 <i>Hz</i> | X | X | 400 <i>Hz</i> – 4 <i>kHz</i> |
| Transitional Sounds | X | X | X | 0 – 4 <i>kHz</i> |
| Transients | X | X | X | 0 – 4 <i>kHz</i> |
| Voiced Plosives | X | 0 – 500 <i>Hz</i> | X | 500 <i>Hz</i> – 4 <i>kHz</i> |
| Voiced Fricatives | X | 0 – 2 <i>kHz</i> | X | 2 <i>kHz</i> – 4 <i>kHz</i> |
| Silence | X | X | 0 – 4 <i>kHz</i> | X |

stable regions of speech such as stable portions of vowels. Transients are defined in the next paragraph.

The segmentation algorithm works on a sample basis as opposed to the enhancement algorithm that works on a frame basis. Therefore, in the enhancement phase, a speech frame may include speech samples from more than one sound. If the speech frame includes samples from two sound classes, the system picks the sound class that has a higher portion of the speech frame. If more than two classes are included, then the frame is labeled as transient.

6.3.2 The MMSE Estimator

The gain function of the estimator is modified by taking into account the risks of misestimation based on the algorithm proposed in [15]. The modified gain function is

$$G' = G^{P_s} G_{min}^{1-P_s} \quad (46)$$

where G is the original gain function of the MMSE estimator discussed in Chapter 4, G_{min} is set to 10^{-3} , and P_s is the signal presence probability. The term signal presence probability is used to indicate the probability of existence of a perceptually important speech signal. The spectral bands that have a high chance of carrying useful speech information are mildly suppressed while low probability bands are more severely suppressed.

6.4 Methods Used in Signal Presence Detection

Three methods are employed in the proposed system to minimize the perceptual costs of misestimations in the enhanced speech. The first method uses the acoustic-phonetic knowledge of speech to oversuppress the perceptually irrelevant parts of the spectrum while preserving the salient cues. Oversuppressing the irrelevant parts of a consonant spectrum does not generate a significant distortion in perceptual quality. Moreover, it increases the intelligibility in noisy environments because the residual noise in the irrelevant parts of the spectrum can degrade intelligibility. For example, in unvoiced fricatives, such as /s/, the speech signal is typically concentrated above 2 kHz. Therefore, P_s for frequencies below 2 kHz can be set to 0. Fig. 14 shows a case where preserving the signal at the higher frequencies while oversuppressing the lower frequencies helps distinguish an unvoiced fricative from a voiced fricative. This method not only suppresses a significant amount of noise in the signal, but also increases the DRT scores for sibilation as will be shown in Table 12.

In the voiced speech spectrum, most of the speech energy is concentrated in high-signal-power (HSP) locations. The signal presence detection system must detect those high-signal-power (HSP) locations to generate high quality speech, which is not always possible with the *SNR* methods. The second method utilizes the GEMS signal for detecting (HSP) locations in the voiced speech spectrum. HSP locations detected by the GEMS signal are assigned a signal presence probability (P_s) of 1 while non-HSP frequencies are assigned a P_s of 0. The HSP detector described in [29] is used. An illustration of HSP detection is shown in Fig. 7.

The third method uses estimated *SNR* at each frequency and assigns soft P_s values as proposed in [15]. This method is used only when silence is detected because of the problems with the *SNR*-based signal presence detection mentioned in the previous chapter.

A different strategy is used for each sound class based on the acoustic-phonetic characteristics of the sound classes and knowledge of human perception using the combination of the three methods. The proposed strategies are described in the next section.

6.5 Sound Class Dependent Signal Presence Detection

The proposed system combines the segmentation information, HSP locations, and the estimated SNR to maximize the intelligibility of speech sounds using a signal presence detection framework. The key idea is to oversuppress the noisy signal to improve the quality without significantly distorting the perceptual cues that are vital for identifying the speech sounds.

The algorithm segments noisy speech into vocalics, unvoiced fricatives, unvoiced plosives, transitional sounds, transients, voiced plosives, voiced fricatives, and silence as described in Section 6.3. The signal presence detection strategy for each sound class is discussed below, and a summary is shown in Table 10.

6.5.1 Vocalics

Vocalics are important determiners of the perceptual quality of speech [58], and their quality primarily depends on the formant locations and bandwidths that are mainly determined by the HSP locations. Therefore, the HSP detector is used for the vocalic class, and P_s is set to 1 for the HSP locations.

Undersuppression of non-HSP locations can create noticeable residual noise [24],[130], and oversuppression of the non-harmonics emphasizes the harmonic trajectories in the spectrogram which can be important for the intelligibility of a neighboring consonant [25]. For example, formant transitions are critically important for identifying the initial plosive sounds. Therefore, P_s is set to 0 for the non-HSP locations.

6.5.2 Unvoiced Fricatives

Acoustic-phonetic knowledge can be used for sibilated unvoiced fricatives such as /s/ and /sh/. The spectral cues for these sounds are typically above 2 kHz , and they do not contain significant speech energy in the lower frequencies [46]. Therefore, P_s is set to 0 for the 0 – 2 kHz for these sounds.

Unsibilated fricatives, such as /th/ and /f/, have significant energy below 2 kHz [46]. However, oversuppression of the spectrum below 2 kHz improves intelligibility by emphasizing the high frequency frication, and it does not create significant confusion between

sibilated and non-sibilated unvoiced fricatives, because of the loss of low-frequency information, as discussed in the experiments section.

The *SNR* information is available above 2 kHz for oversuppressing the noisy signal. However, above 2 kHz , using the *SNR* method runs the risk of removing the low-energy frication cues. Moreover, the signal is expected to be everywhere in the spectrum above 2 kHz for unvoiced fricatives. Thus, P_s is set to 1 for the $2\text{ kHz} - 4\text{ kHz}$ range.

6.5.3 Unvoiced Plosives

Typically, the fundamental frequency of voiced sounds is in the $0 - 400\text{ Hz}$ range, and first couple of harmonics in that range are important in pitch perception [101]. Therefore, the $0 - 400\text{ Hz}$ range is assigned a P_s of 0, and the low-frequency parts of the spectrum are severely suppressed to decrease the chance of confusing an unvoiced plosive with a voiced sound [26]. Moreover, the burst cues of plosives are usually not at these low frequencies, and the risk of erasing perceptually important burst cues is low.

P_s is set to 1 for the $400 - 4000\text{ Hz}$ range of unvoiced plosives. The *SNR* method does not perform well for this region since low-*SNR* burst cues may be erased by the algorithm, and *SNR* estimation errors can further create significant loss of cues or generate false cues.

6.5.4 Voiced Plosives

First two or three harmonics are important in the perception of voicing as described above. The *SNR* method is not always capable of resolving the fundamental frequency at low *SNRs*. In Fig. 13, a spectrum of a voiced plosive is contaminated with M2 noise. Not only is the harmonic pattern in the low frequencies are distorted, but also the *SNR* information is not useful in oversuppressing the non-harmonics, since all frequencies below 500 Hz have high *SNRs*.

The GEMS signal can accurately indicate the first few harmonics for voiced plosives. Hence, the HSP method is used for the first 500 Hz of voiced plosives to emphasize the harmonics at these low frequencies. Similar to the unvoiced plosives case, a P_s of 1 is used for all frequencies above 500 Hz .

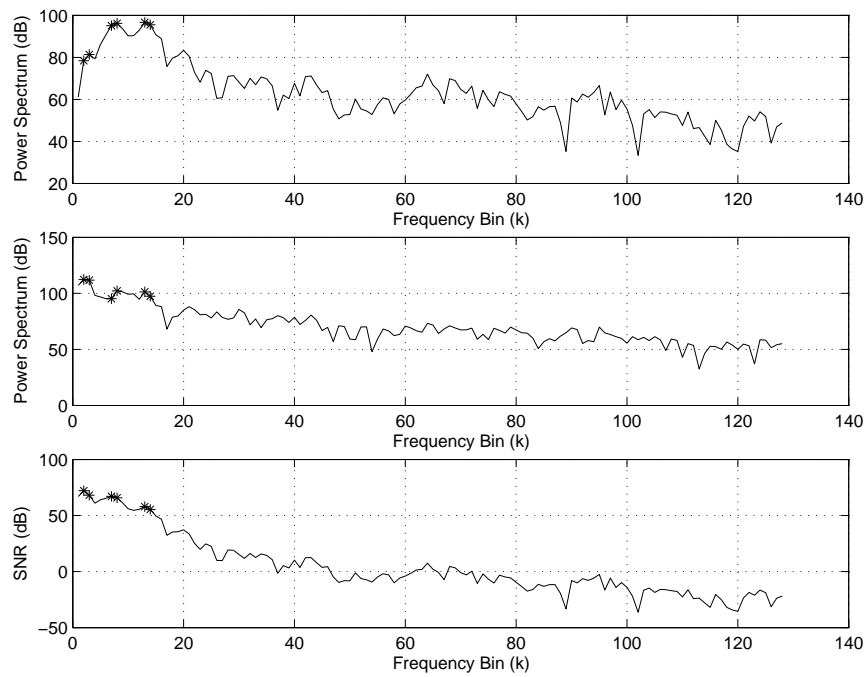


Figure 13: Power spectrum of a 20 msec speech sample taken from the bust location of a clean plosive sound is shown in the top figure. The middle figure shows the same spectrum contaminated with M2 tank noise at 0 dB. The bottom figure shows the SNR (dB) of the noisy spectrum. Mean noise power is used to estimate the SNR. HSP locations in the first 500 Hz of the spectrum are marked.

6.5.5 Voiced Fricatives

In voiced fricatives, both lower and higher frequencies contain important cues. Low frequencies are important because they indicate voicing in the signal, and high frequencies are important because they indicate frication in the sound.

The HSP method is used for the frequency range of $0 - 2 \text{ kHz}$ to create a harmonic pattern in this range to perceptually emphasize voicing. Although a fricative sound may not have a harmonic structure up to 2 kHz , using the HSP method is not found to create any perceptually audible artifact. Similar to unvoiced fricatives, the P_s is set to 1 for the $2 \text{ kHz} - 4 \text{ kHz}$ range to preserve the low-energy frication cues.

6.5.6 Transitional Sounds and Transients

Although the GEMS device can often receive a signal during transitional sounds and transients, the HSP detector sometimes fails to detect perceptually important HSP locations because of the relatively weaker GEMS signal, particularly at the voicing onsets. Therefore, the HSP detector is not used for these voiced sounds.

Low energy sounds, such as nasals, are prone to suppression of valuable cues with the *SNR* method. Furthermore, the human ear can mask a significant amount of background noise at the rapidly changing transitional regions [118]. Therefore, the *SNR* method is also not considered for the transitional sounds and transients.

Given the limitations of the HSP method and the *SNR* method, P_s is set to 1 for the whole spectrum for transitional sounds and transients.

6.5.7 Silence

Undersuppression has a high risk of creating residual noise and tonal artifacts for silence segments. Therefore, the signal is oversuppressed at all frequencies when silence is detected. The acoustic-phonetic knowledge can be used to completely suppress the silence regions. However, listeners prefer a low-level comfort noise, and enhanced speech sounds less pleasing when silence segments are completely suppressed. Moreover, the segmentation algorithm is not perfect, and the *SNR* method reduces the risk of oversuppressing the speech cues

Table 11: Parameter values used by the proposed speech enhancement system.

| Parameter | Value |
|---------------|--------------|
| Frame Length | 256 samples |
| Frame Rate | 128 samples |
| Window Type | Hanning |
| G_{min} | 10^{-3} |
| Sampling Rate | 8 <i>kHz</i> |

in the case of a segmentation error, since the high *SNR* speech cues can be detected and preserved with the *SNR* method. Therefore, the *SNR* method is used to detect the signal for the whole spectrum.

6.6 Experiments

The proposed system is used at the front-end of the 2.4 kbps MELP speech coder and compared with the MELPe NATO speech coding standard. The MELPe coder uses the MMSE estimator (M-MMSE) described in [89] in concatenation with the MELP speech coder. The Diagnostic Rhyme Test (DRT) is used to measure the intelligibility of coded speech. Experiments are conducted as per ITU recommendations [4]. The subjects consisted of six native English speakers, and each subject had at least one hour of previous experience with the DRT testing. Test materials were roughly 15 minutes duration, containing 696 DRT syllables spoken by three male and three female, trained native English speakers.

M2 noise was used for testing since it had been found to be the most challenging test environment used during the development of MELP coders. Speech is recorded using a noise-cancelling microphone, with *SNR* approximately 0 *dB*¹. Parameters of the speech enhancement system are given in Table 11.

Description of the DRT is made in the next section. Experimental results are presented and discussed in Section 6.6.3.

¹The speech and sensor data is from a subset of the ARCON speech database that is created as part of the DARPA Advanced Speech Encoding (ASE) program

6.6.1 Description of the Diagnostic Rhyme Test (DRT)

The DRT is an intelligibility test that for diagnosing weaknesses in a voice communication system [140]. Consonant-Vowel-Consonant (CVC) monosyllables are presented to a set of listeners continuously; each listener makes a choice between the true syllable and an alternative false syllable. The true and false syllables differ only in their initial consonants with these differing only by one distinctive feature. Thus, the test is designed to diagnose intelligibility problems of initial consonants. It is well known that this test correlates highly with other, more complex, intelligibility tests.

There are six distinctive features that are used in DRT, which include voicing, nasality, compactness, sibilation, sustention, and graveness. They are briefly described in Section 6.6.3. After a complete test is administered to a given listener, intelligibility scores for each feature and total intelligibility score is calculated. The intelligibility score for a feature i is calculated using

$$Score(i) = \frac{(Number\ of\ correct\ answers)}{(Total\ number\ of\ questions)} 100, \quad (47)$$

and

$$Total\ Score = \frac{1}{6} \sum_{i=0}^6 Score(i). \quad (48)$$

The score for each individual feature can be used to diagnose the particular problems with the communication channel. All listeners' scores are used for calculating averages and standard errors, which are then used for significance testing.

DRT results of two systems are shown in Table 12. These will be discussed in Section 6.6.3.

6.6.2 Analysis of Segmentation Errors

In the proposed system, errors in segmentation can potentially degrade the DRT scores. Therefore, segmentation errors are analyzed to understand their effect in the final DRT scores and results are shown in Table 13. Long-duration, high-energy voiced sounds can be detected by the segmentation algorithm because the p-mic and the GEMS signals can reliably indicate these sounds. However, problems occur frequently with fricatives and

Table 12: The DRT scores for each distinctive feature compared for MELP encoded speech in 0 dB M2 noise environment.

| Feature | M-MMSE | Proposed System | p value |
|-------------|--------|-----------------|---------|
| Voicing | 68.62 | 75.63 | 0.12 |
| Nasality | 60.7 | 91.86 | 0.0001 |
| Sustention | 60.3 | 60.09 | 0.19 |
| Sibilation | 74.41 | 81.79 | 0.18 |
| Graveness | 68.27 | 74.43 | 0.14 |
| Compactness | 77.09 | 77.5 | 0.21 |
| Total | 68.23 | 76.88 | 0.0001 |

plosives. For example, voiced plosives and voiced fricatives were rarely detected as shown in Table 13. Therefore, the analysis of segmentation errors is constrained to plosives and fricatives.

The following procedure was used for analyzing the segmentation errors. Spectrograms and segmentation data for all monosyllables were aligned and analyzed manually for errors. An example alignment is shown in Fig. 14. Two types of errors are considered. A detection error occurs when the correct consonant is not detected in its time boundaries. A confusion error occurs when the consonant is confused with another sound.

The detection and confusion information is recorded for the initial consonants only, as described above. The segmentation algorithm sometimes detects more than one initial consonant. The time boundaries for the correct sounds are not known since the clean signal is not available. Therefore, subjective judgment is used to decide which sound is detected at the time boundary of the correct sound. Ambiguous cases rarely occurred, and errors in the subjective judgments were not expected to create a statistically significant bias.

The sound detection rate for four consonant types are shown in Table 13. Confusion rates between the four consonant types are shown in Table 14. The detection and confusion data is used for the discussion of the results below.

6.6.3 Results

Experimental results are shown in Table 12. The proposed system has a higher mean DRT score than the M-MMSE system. Statistical significance of the results are measured using paired student's t-test with 95 percent confidence interval. The p value indicates

Table 13: Detection rates (%) of the four consonant sounds using the segmentation algorithm.

| | Detected | Not Detected |
|---------------------|----------|--------------|
| Unvoiced Plosives | 53 | 47 |
| Unvoiced Fricatives | 43 | 57 |
| Voiced Fricatives | 4 | 96 |
| Voiced Plosives | 10 | 90 |

Table 14: Mutual Confusion rates (%) among four consonant classes using the segmentation algorithm.

| | Voiced Plosives | Unvoiced Plosives | Voiced Fricatives | Unvoiced Fricatives |
|---------------------|-----------------|-------------------|-------------------|---------------------|
| Voiced Plosives | X | 38 | 0 | 4 |
| Unvoiced Plosives | 0 | X | 0 | 29 |
| Unvoiced Fricatives | 3 | 4 | 0 | X |
| Voiced Fricatives | 19 | 4 | X | 8 |

the probability that the proposed system has a higher score than the M-MMSE system. Analysis of results for each feature type is described below.

6.6.3.1 Voicing

The voicing feature discriminates a voiced consonant from its unvoiced cognate. For example, voiced plosives and unvoiced plosives can be discriminated using the voicing feature. The proposed system improves the voicing feature compared to the M-MMSE system as shown in Table 12.

One of the expected effects of using the GEMS signal was improvement in the voicing feature. The idea was to detect the voiced sounds using the segmentation algorithm, and emphasize harmonicity and voicing using the HSP detector. However, the improvement in the voicing feature is modest and below the expected gains. The major reason that limited the improvement in the voicing feature was found to be the errors in the segmentation algorithm as discussed below.

Since the auxiliary sensors can indicate voicing in speech, when confusion occurs, voiced sounds are expected to be confused with other voiced sounds, and unvoiced sounds are expected to be confused with silence or other unvoiced sounds. However, this is sometimes not the case, particularly for voiced plosives and voiced fricatives as shown in Table 14.

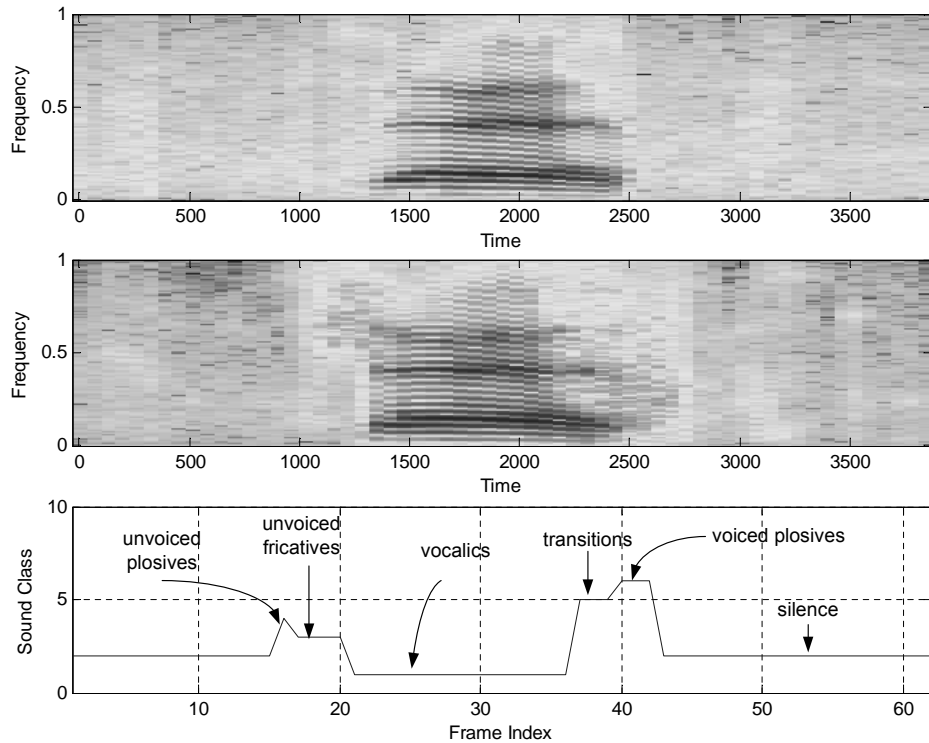


Figure 14: Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the voicing feature. The DRT word pairs are “SAID” and “ZED,” and the true word is “SAID.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

Confusions of voiced sounds with unvoiced sounds can be attributed to the fact that the auxiliary sensor signals are also noisy, and they typically have lower energy for the voiced fricatives and voiced plosives compared to the high-energy voiced sounds such as vowels. Such confusions limit the improvement in the voicing feature.

Most of the voiced plosives and voiced fricatives are not detected by the segmentation algorithm. They are typically labeled as transition sounds when not confused with an unvoiced sound. The transition sounds are enhanced without any oversuppression. Therefore, the low detection rate of voiced consonants did not significantly degrade the voicing feature.

Voiced plosives are frequently confused with unvoiced plosives as shown in Table 14. However, this usually does not affect the voicing feature because the confusions are found to occur only at the burst location but not at the following voiced aspiration phase. Thus, the burst is labeled as unvoiced plosive but the following aspiration part is typically labeled

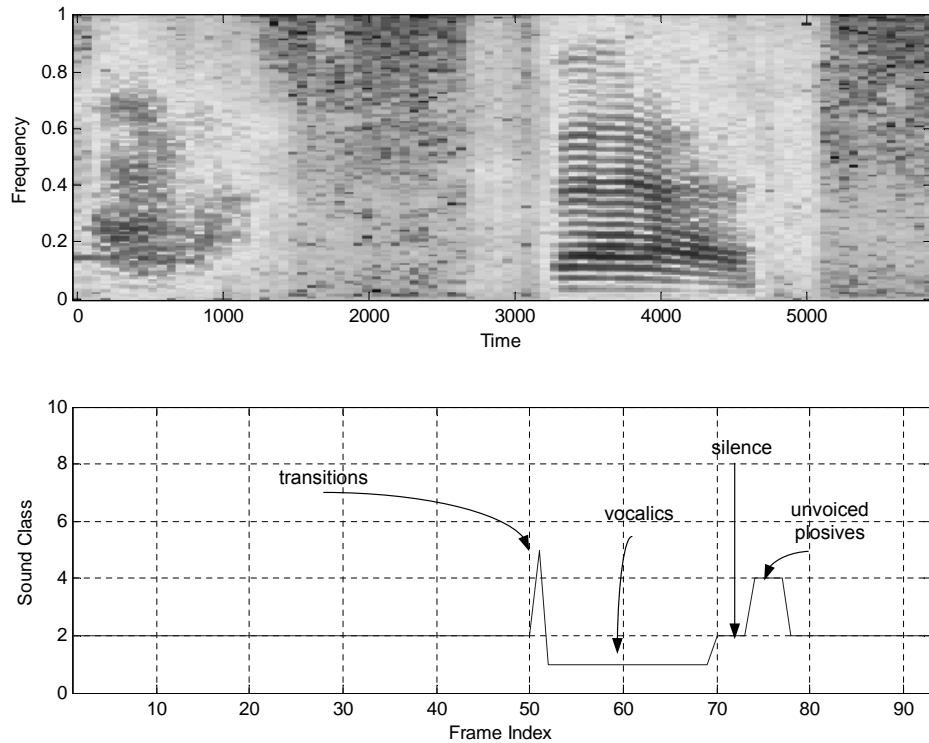


Figure 15: An example case is shown where the segmentation error causes confusion in voicing. The correct word is “VAST,” but the enhanced word sounds like “FAST,” because the cues related to voicing and frication are suppressed.

with a voiced sound. Therefore, significant amount of voicing cues could be preserved for these cases.

Detection rates are low for unvoiced plosives and unvoiced fricatives as shown in Table 13. However, confusion of unvoiced plosives and unvoiced fricatives with voiced sounds happen rarely. Therefore, low detection rates for these sounds do not create major confusion in the voicing feature.

Some of the cases which were consistently confused by most of the listeners in testing the voicing feature were analyzed. Almost all of those cases were related to the fricatives. Undetected fricatives, which were confused with silence, are found to seriously degrade the performance. Fig. 15 shows a case where the fricative is not detected at all. This caused the word “VAST” to sound like “FAST”, and an error in identifying the voicing feature occurred. Similar other cases were observed where the problems with undetected fricatives degrade the voicing score.

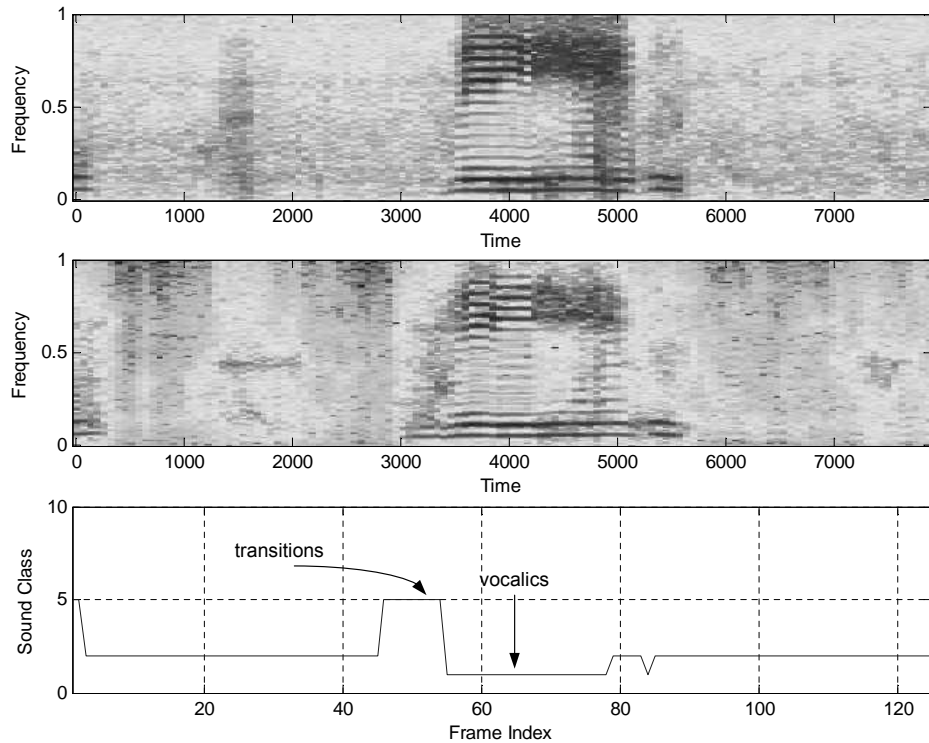


Figure 16: Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. The DRT word pairs are “NEED” and “DEED,” and the true word is “NEED.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

Similar to the case of degradations, improvements in the voicing feature were also observed mostly with the fricatives. Figure 14 shows a case where the segmentation algorithm preserved the cues related to voicing. The segmentation algorithm indicated an unvoiced fricative, and the enhancement system preserved the unvoiced fricative cues whereas the M-MMSE system removed almost all the cues related to the unvoiced fricatives. Listeners preferred “ZED” for the M-MMSE system while they preferred “SAID” for the proposed system.

Although errors in the segmentation algorithm degrades the performance, it also helps improve intelligibility in some cases as discussed above. As a result, moderate improvement in the voicing feature compared to the M-MMSE system is obtained using the proposed system.

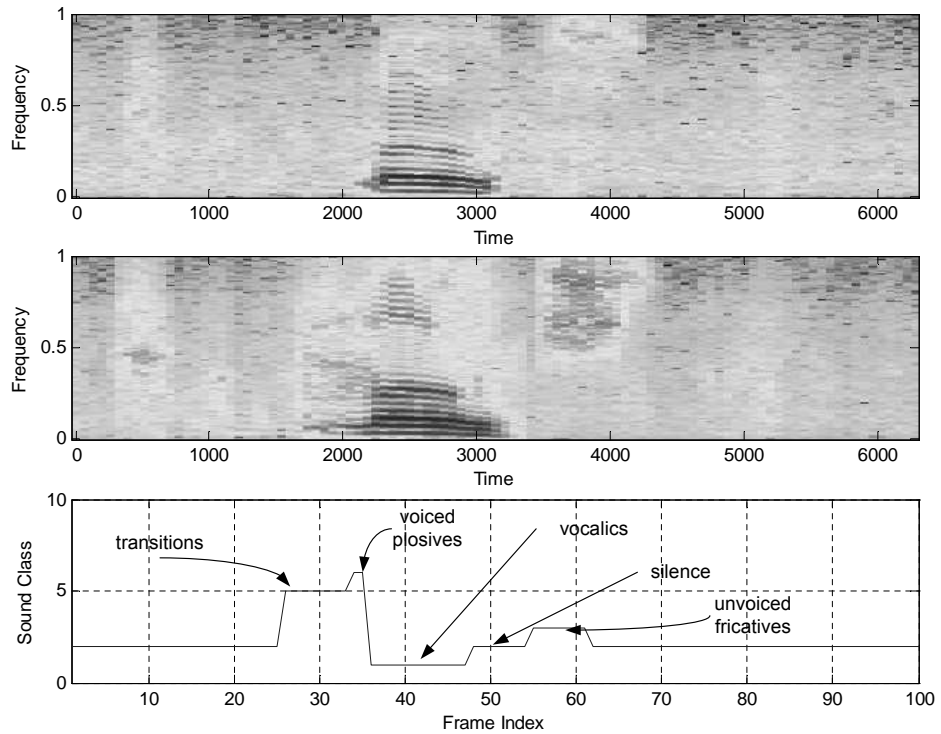


Figure 17: Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. The DRT word pairs are “MOOT” and “BOOT,” and the true word is “MOOT.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

6.6.3.2 Nasality

Nasality feature discriminates between the nasalized consonants and the oral consonants. For example, nasals can be distinguished from vowels using the nasality feature. The proposed system has a substantially higher nasality score compared to the MMSE system.

The nasality feature, as measured by the DRT, was improved compared to the M-MMSE estimator. The M-MMSE system was found to erase the low-energy nasality cues frequently which create confusions between ($/m,n/$) and ($/b,d/$) sounds as shown in Figures 16, 17, 18, 19. The proposed system can successfully preserve those low-energy cues because the segmentation algorithm typically detects a transition sound, which is enhanced without any oversuppression, during nasals.

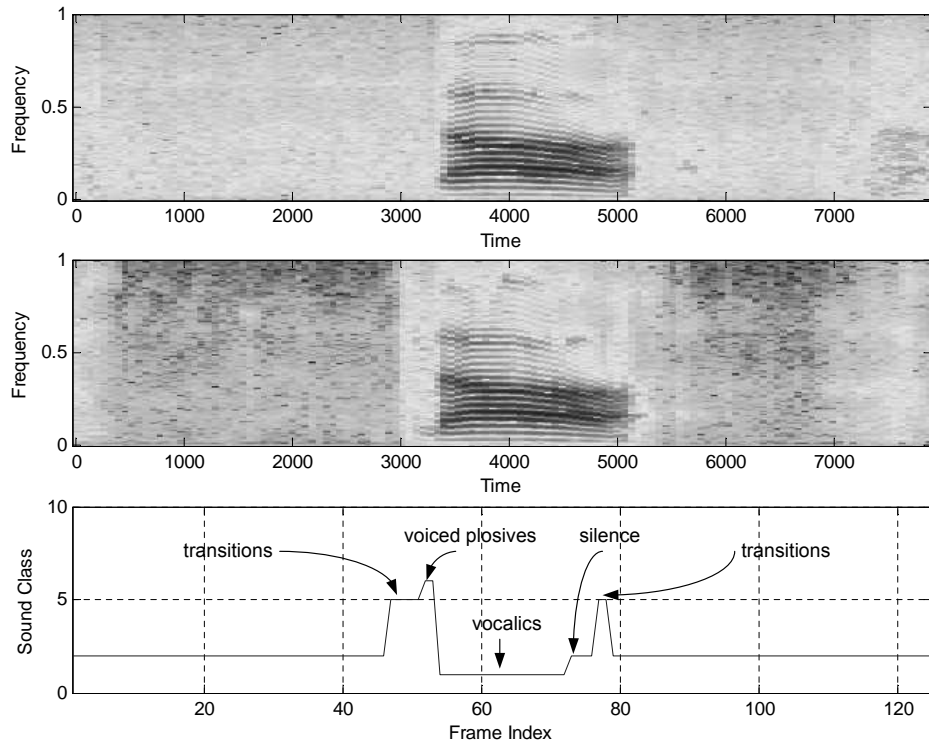


Figure 18: Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in the nasality feature. The DRT word pairs are “GNAW” and “DAW,” and the true word is “GNAW.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

6.6.3.3 Sibilant

The sibilant feature is used to distinguish between irregular and regular waveforms. For example, the sibilant feature can be used to distinguish between an unvoiced fricative and unvoiced plosive. The proposed system improves the sibilant feature compared to the M-MMSE system.

The improvement in the sibilant feature is limited mainly because the segmentation algorithm can sometimes fail to detect the fricative sound which cause a deletion of the sibilant cues as shown in Fig. 15. Similar to the voicing feature, in some cases, the segmentation information helped with the sibilant feature. A sample case is shown in Fig. 20. This example is interesting because voicing could not be detected in the initial consonant, but frication was detected. Therefore, although the initial consonant sounds

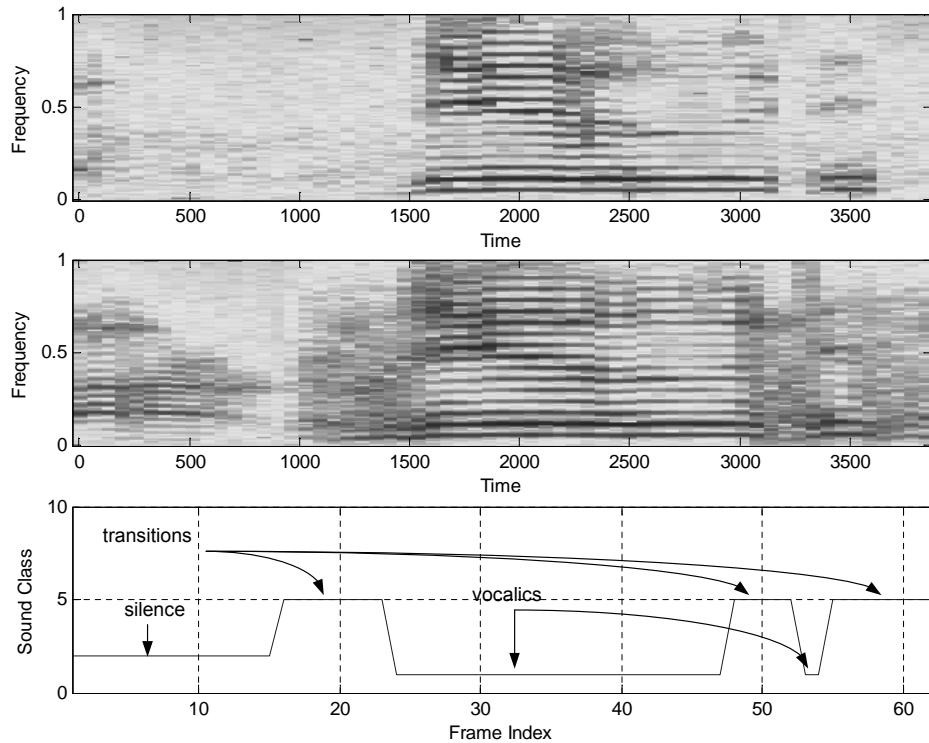


Figure 19: Spectrograms of the M-MMSE algorithm and the proposed system are compared for a monosyllable sample to show the improvement in nasality. The DRT word pairs are “NUDE” and “DUDE,” and the true word is “NUDE.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

like the unvoiced fricative /s/, listeners could still identify the correct monosyllable.

6.6.3.4 Graveness

Graveness feature is used to detect the ratio of the high frequency energy to the low frequency energy in a speech sound. Graveness of a consonant increases as its energy concentrates more at the lower frequencies. For example, graveness can be used to distinguish between “WEED” and “REED”. The speech energy is concentrated at the higher frequencies for the /w/ sound compared to the /r/ sound. The proposed system improves the graveness feature compared to the M-MMSE system.

The improvement in the graveness feature could not be attributed to any one aspect of the proposed system. This result is partly related to the fact that, as opposed to voicing, nasality, and sibilation features, the graveness feature is related to a comparison of a variety

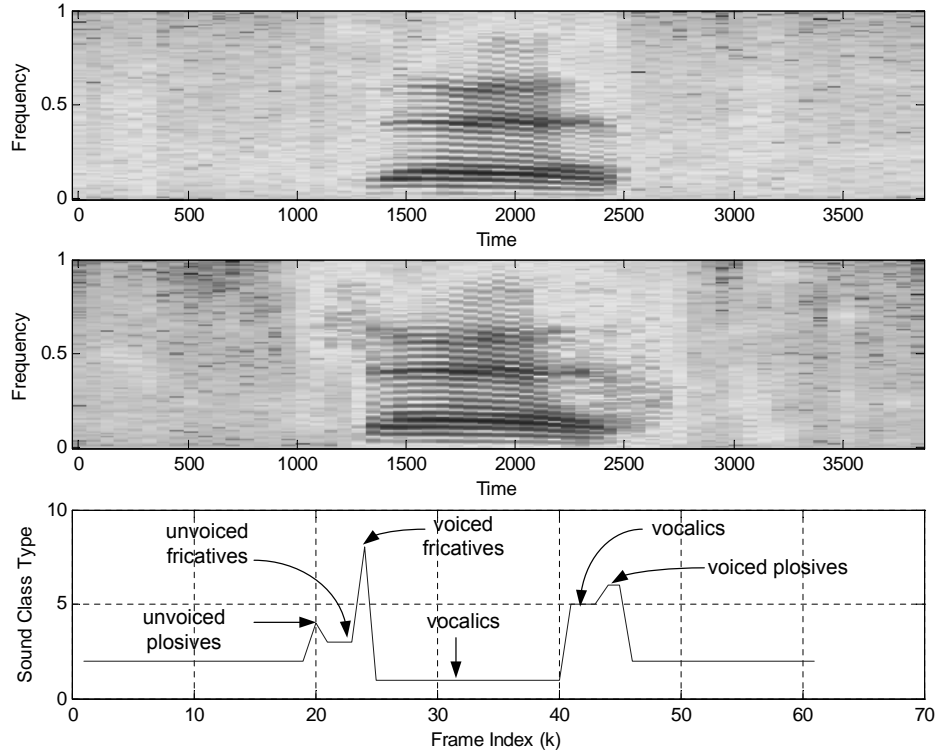


Figure 20: Spectrograms of the M-MMSE algorithm and the proposed system are compared to show the improvement in sibilation for a monosyllable sample. The DRT word pairs are “THEE” and “ZEE,” and the true word is “ZEE.” The top figure illustrates the output of the M-MMSE estimator, the middle figure illustrates the output of the proposed system, and the bottom figure illustrates the segmentation data.

of sound classes.

6.6.3.5 Sustention

The sustention feature is used to indicate if the consonant is sustained or interrupted. For example, plosives are interrupted sounds while fricatives are sustained. The sustention feature has similar scores for both systems.

Plosive sounds are short-duration events that change faster than the time resolution of the MELP coder [139]. Therefore, even for clean speech, the sustention feature, which is related to discriminating between plosives and fricatives, is low compared to other features [144]. This is partly the reason why the sustention feature did not change significantly.

6.6.3.6 Compactness Feature

The compactness feature is used to detect if the energy of the consonant is concentrated at the mid frequencies. For example, compactness can be used to distinguish between “KEY” and “TEA”. Energy of the /k/ sound is more concentrated at the mid frequencies. Hence, it is more compact compared to the /t/ sound. The compactness feature has similar scores for both systems.

6.7 Summary and Conclusions

A multi-sensor speech enhancement system is proposed that uses sound-class dependent noise suppression, where each sound segment is enhanced based on its acoustic-phonetic properties. Acoustic phonetic knowledge of speech, HSP location information for voiced speech, and *SNR* information is combined in a speech presence detection framework to minimize the risk of erasing the perceptually important speech cues. Several observations are made using the experiment results:

- Nasality feature improved substantially while voicing, sibilant, and graveness features improved modestly. The DRT scores improved from 68.62 to 91.86 for the nasality feature.
- The improvement in nasality is found to be related with the low performance of the speech presence detection algorithm employed in the M-MMSE algorithm. The segmentation algorithm can detect nasal sound successfully, and low-energy nasality cues are preserved in the proposed system.
- The improvement in voicing and sibilant features found to be mostly related to fricatives. The M-MMSE algorithm fails to detect the low-energy sibilant features. The segmentation algorithm also fails to detect sibilant and/or voicing for some cases. However, the proposed system could improve the DRT scores by 7 points for voicing and sibilant features.
- Confusion of plosives with other sounds or silence did not generate significant degradation in performance in any of the features. Correct detection of plosives also does

not seem to help improve the intelligibility of plosive sounds.

- The sustention feature which is designed to distinguish between plosives and fricatives did not change significantly partly because the time-resolution of the MELP coder is not high-enough to resolve the rapid spectral transitions during plosive sounds.
- The improvement in graveness feature could not be attributed to any specific aspect of the MMSE system or the proposed system.
- A significant change is not observed with the compactness feature.

CHAPTER VII

USING A SIMULATED VOICING FEATURE FOR NOISE-ROBUST ASR

Since the speech enhancement system proposed in Chapter 6 uses an automatic segmentation/classification block, possible use of GEMS signal for a noise-robust segmentation/classification task is further investigated using HMM-based automatic speech recognition (ASR) techniques. A general background of the noise-robust ASR field is given in Chapter 2, and some of the noise-robust ASR systems that use auxiliary sensors are discussed below.

Auxiliary sensors, such as the GEMS device, bone microphone, or physiological microphone, can provide reliable information for noise-robust ASR applications. One way of exploiting this information is using the auxiliary sensor signal to enhance the speech signal [53],[147] since the noise-robust sensor signal can help design better speech enhancement systems. Speech/non-speech detection using a bone microphone is also done in [147]. In this work, a voicing feature is extracted from the GEMS signal for increased noise immunity in an ASR task. Voicing is defined as the energy of the GEMS signal in the 0-500 Hz frequency band.

A voicing feature has been used in previous systems which mostly focus on clean speech. In [151], three different voicing detection algorithms are compared, and the voicing feature is used for small and large vocabulary tasks. In [132], an autocorrelation based voicing measure is used, and MFCC features are fused with the voicing feature and the first and second order derivatives of the voicing feature. In [86], fundamental frequency and voicing are used with the MFCC features using Linear Discriminant Analysis (LDA).

The fundamental problem in the systems mentioned above is the difficulty in extracting the voicing feature reliably at low SNRs. Moreover, the problem becomes even more severe under non-stationary noise conditions. Misclassifications in voicing detection can decrease

ASR performance rather than increase it at low SNRs. Therefore, only marginal improvement in word-error-rate (WER) could be obtained with the previous systems that use the voicing feature.

This chapter addresses the problem of efficient utilization of a reliable voicing feature extracted from an auxiliary sensor for noise-robust ASR systems. The first contribution is two different strategies for deploying the voicing feature for increased noise robustness in the ASR system. It is shown that both strategies can substantially increase the noise robustness using a voicing feature compared to the baseline system, the ISIP ASR engine developed at Mississippi State University. The second contribution is a proposed system that uses a combination of the two techniques to efficiently employ the voicing feature at all SNRs. It is shown that the two proposed strategies are complementary to each other, and when used together, substantial performance improvements can be obtained at all noise environments and all SNRs.

The first strategy described in this work is the feature fusion (FF) method. The voicing feature is directly fused with the speech-based features and recognition is done using the final feature vector. The FF method is found to perform well at all SNR levels while it is most effective at 10 dB for all investigated noise types.

The second strategy is a data marginalization (DM) method where recognition is performed using only the voicing feature for low-energy speech frames while the feature fusion method is used for high-energy speech frames. The rationale behind this approach is the fact that low-energy speech segments are typically significantly more sensitive to background noise compared to high-energy segments. Indeed, experiment results show a substantial performance improvement below 15 dB with the data marginalization method.

Experiment results show that the two strategies are complementary to each other. The FF method performs better than the DM method down to 10 dB while the DM method performs better than the FF method below 10 dB. The proposed system uses a global SNR based switch system that, for a given speech utterance, uses the FF method if the global SNR is above 5 dB, and the DM method is used if the global SNR is below 10 dB to efficiently utilize the voicing feature. Experiment results with the Aurora2 database show

substantial improvement over the baseline system at all SNRs.

This chapter is organized as follows. The two methods for utilizing the voicing feature as well as the proposed system that uses both methods are described in Section 7.1. The algorithm for voicing detection and high-energy speech detection are also described in Section 7.1. The auxiliary sensor signal is not available for any standard ASR database. Therefore, a simulation methodology is used to simulate the voicing feature and high-energy speech detection for the Aurora2 database. The simulation methodology for the ASR experiments are described in Section 7.2. Experiment results are discussed in Section 7.3.

7.1 Three Strategies for Utilizing the Voicing Feature

7.1.1 Feature Fusion Method

The voicing feature can be directly fused with the MFCC features and fed into the ASR engine. This approach is called the feature fusion (FF) method in this work. Although it is a straightforward way of utilizing the voicing feature, it is found to perform well particularly between 5 dB and 15 dB as shown in Section 7.3.

The problem that is observed with the FF method is that below 10 dB the performance improvement rapidly starts decreasing. Although a reliable estimate of the voicing feature increases the overall performance, the distortions in the speech features significantly degrade the system performance below 10 dB which is discussed below.

7.1.2 Analysis of the SNR-Dependent Performance Gain with the Voicing Feature Method

In a typical Hidden Markov Model (HMM) based ASR system, log-likelihood scores (L) for many possible word strings are evaluated. The word string with maximum L is output at the end of the decoding process. Since L is the distance metric used in decoding, evaluation of L is explained below to analyze the limitation of the feature fusion method.

Log-likelihood score of HMM-state k at frame t is

$$L(k, t) = \log \left[\sum_{j=1}^K w_{kj} N(o_t; \mu_{kj}, C_{kj}) \right] \quad (49)$$

where K is the number of Gaussian mixtures, o_t is the observation vector at frame t , w_{kj}

is the mixture weight, C_{kj} is the covariance matrix, and μ_{kj} is mean vector of the j^{th} mixture component of state k . In practice, for efficient calculation, the log-likelihood score is approximated by

$$L(k, t) \approx \max_j \{ \log(w_{kj}) + \log(N(o_t; \mu_{kj}, C_{kj})) \}. \quad (50)$$

where $\log(w_{kj})$ is a constant. In case of a diagonal covariance matrix,

$$\log(N(o_t; \mu_{kj}, C_{kj})) = \sum_{i=1}^M \log [N(o_t(i); \mu_{kj}(i), C_{kj}(i, i))] \quad (51)$$

where M is total number of features, $o_t(i)$ is the i^{th} element of the observation vector, $C_{kj}(i, i)$ is the i^{th} diagonal element of the covariance matrix, and $\mu_{kj}(i)$ is the i^{th} element of the mean vector.

If the MFCC features derived from the speech signal are fused with the voicing feature derived from the sensor signal, then Eq. 51 can be divided into two components

$$\log(N(o_t; \mu_{kj}, C_{kj})) = L_{MFCC}(k, t) + L_v(k, t) \quad (52)$$

where, assuming the voicing feature is fused to the end of the MFCC features,

$$L_{MFCC}(k, t) = \sum_{i=1}^{M-1} \log [N(o_t(i); \mu_{kj}(i), C_{kj}(i, i))] \quad (53)$$

and

$$L_v(k, t) = \log [N(o_t(i); \mu_{kj}(i), C_{kj}(M, M))] \quad (54)$$

For noisy speech,

$$\log(N(o_t; \mu_{kj}, C_{kj})) = L_{MFCC}(k, t) + L_v(k, t) + L_d(k, t) \quad (55)$$

where L_d term compensates for the distortion in the MFCC features. The distortion ratio can be defined as

$$D_1 = \frac{L_d(k, t)}{L_{MFCC}(k, t) + L_v(k, t)}. \quad (56)$$

If the voicing feature is not used, and the log-likelihood evaluation is based on the MFCC features only, then

$$\log(N(o_t; \mu_{kj}, C_{kj})) = L_{MFCC}(k, t) + L_d(k, t), \quad (57)$$

and the distortion ratio is

$$D_2 = \frac{L_d(k, t)}{L_{MFCC}(k, t)}. \quad (58)$$

The log-likelihood score $L_v(k, t)$ contributed by the voicing feature helps decrease the distortion ratio as shown in Equations 57 and 58. However, with decreasing SNR, expected value of L_d increases. Therefore, D_1 approaches D_2 , and the voicing information becomes less effective with decreasing SNR.

To avoid this problem and efficiently utilize the voicing feature, a data marginalization method is proposed in the next section.

7.1.3 Data Marginalization Method

Low-energy speech sounds are particularly more susceptible to ambient noise compared to high-energy speech sounds since low-energy sounds can be relatively easily masked by additive noise as shown in Fig. 21. In order to address the noise sensitivity problem of the low-energy speech sounds, only the voicing feature is used when such a low-energy sound is detected. Since the acoustic models contain all the features for all states, a data marginalization (DM) method is used that calculates the likelihood using the voicing feature only if the speech sound is classified as low-energy. After marginalizing the likelihood function over all the MFCC features, for a diagonal covariance matrix, the likelihood function becomes

$$L_m(k, i) = \log(w_{kj}) + \log(\mathfrak{N}(o_t(M); \mu_{kj}(M), C_{kj}(M, M))) \quad (59)$$

where L_m denotes the marginalized likelihood function, and voicing feature is assumed to be in the M^{th} position of the observation vector. All the features contributed by the noisy speech signal are ignored in this formulation.

One can also design a system that detects subbands that are reliable in the low-energy speech frames and use those subbands for recognition. One of the problems with this approach is the difficulty in detecting the reliable subbands particularly at the SNR levels where the marginalization method is used in this work. A subband-based approach is compared with the proposed method in Section 7.3.

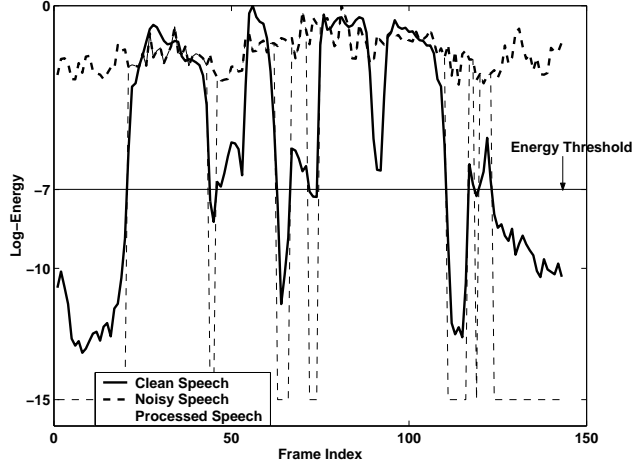


Figure 21: Energy trajectories of the clean speech and the noisy speech with babble noise at -5 dB are compared for an utterance from the Aurora2 database. Masking effect is clear for low-energy regions. Energy trajectory is thresholded at -7 dB to show the high-energy speech segments. The vertical lines indicate the boundaries of the high-energy speech segments.

7.1.4 Proposed System

An overview of the final proposed system is shown in Fig. 22. The DM method is used for speech utterances with global SNRs below a threshold. The FF method is used if the global SNR is above the threshold. The global SNR threshold level is set to 10 dB based on the results presented in Section 7.3. The voicing extraction algorithm is defined in the next section.

7.1.5 Voicing Extraction Algorithm

A hard-decision energy-based algorithm is used to detect the voiced segments using the GEMS signal. Each 20 msec of GEMS frame i is windowed with a Hanning window. Total energy ζ_i in the 100-400 Hz range is calculated for each frame i . This range generally contains at least one harmonic in the case of voicing. Thus, significant energy rise in this band compared to the silence level indicates existence of voicing. A minimum energy level $\zeta_{min,r}$ is set to the mean of energy levels of the first 10 frames in the radar signal, assuming that they are noise-only frames, and a threshold energy level $\zeta_{th,r}$ is set to 6 dB above the

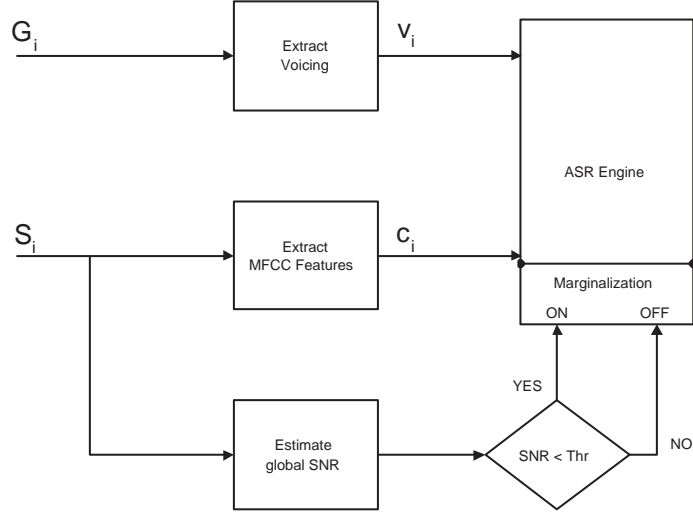


Figure 22: Flow diagram of the proposed system is shown. i^{th} speech frame is denoted with S_i , and i^{th} GEMS frame is denoted with G_i .

$\zeta_{min,r}$. The voicing feature V_i for frame i is

$$V_i = \begin{cases} 1 & \text{if } \zeta_i > \zeta_{th,r}, \\ 0 & \text{if } \zeta_i \leq \zeta_{th,r}. \end{cases}$$

where ζ_i is the energy of the i^{th} frame and r refers to radar.

Although the voicing feature has a discrete value, a Gaussian mixture distribution is used in Eqn. 59 for the voicing feature. The ASR engine has a minimum variance (MV) parameter that avoids numerical problems associated with the small variances of Gaussian distributions. The training program does not allow the variances of Gaussian distributions fall below the MV parameter. Thus, the Gaussian mixture assumption works because the system can model a discrete binary distribution with two Gaussians that have small variances.

7.1.6 High-Energy Speech Detection

Most high-energy speech sounds can be detected using the GEMS signal since there is a strong correlation between voicing and energy in a speech frame. There are two problems that need to be addressed in high-energy speech detection using the GEMS signal. The first problem is determining the energy threshold to define high-energy speech sounds. In

this work, the energy threshold is set based on the performance of the resulting system as discussed below.

Table 15 shows the performance of the DM method with various energy thresholds for energy normalized speech. Energy normalization is performed by normalizing the energy feature of all speech frames with the energy value of the highest energy frame in a speech utterance. Thus, the maximum value of the energy feature is 0 (dB) in the speech utterance after normalization. The highest energy threshold of -5 dB achieves the best performance up to 5 dB SNR. For input SNRs above 10 dB, the information loss because of data marginalization starts degrading the performance, and the optimum threshold decreases rapidly. The threshold of -5 dB is used in the proposed system since it provides the highest performance gain below 10 dB SNR.

The second problem that should be addressed for high-energy speech detection is the difficulty in detecting the speech sounds with energy above -5 dB using the GEMS signal. The energy of the GEMS signal is thresholded similar to the speech signal. The GEMS-energy threshold is tuned to minimize the probability of error at each speech-energy threshold level. P_e values at each threshold level is shown in Table 16. Although P_e is lowest when the speech-energy threshold is -7 (dB), P_e is still low at -5 (dB).

One of the problems with the GEMS device is that there is not enough training data for speech recognition experiments. Therefore, GEMS-based high-energy speech and voicing detection is simulated using the clean speech signal. Details of the simulation methodology for voicing detection and high-energy speech detection are given in the next two sections respectively.

7.2 Simulating Voicing and High-Energy Speech Detection

7.2.1 Simulating the GEMS-based Voicing Detection

Voicing is detected from the clean speech signal using the algorithm described in Section 7.1.5. The threshold level $\zeta_{th,s}$ for the speech signal is set to $\zeta_{min,s} + 8(dB)$ where $\zeta_{min,s}$ is the average energy of the first 10 frames. The first 10 frames in an utterance are assumed to be noise-only frames.

Table 15: Averaged word error rate (WER) results for four different noise types in the Aurora2 database are shown. Data marginalization method is used. Various energy thresholds are investigated, and the performance is compared with the baseline system. The WER results are presented for SNR levels from -5 dB to 20 dB as well as the clean case. The lowest WER at each SNR has bold font.

| SNR/Threshold | -5 dB | -7 dB | -9 dB | -11 dB | Baseline |
|---------------|--------------|--------------|-------------|--------|-------------|
| -5 dB | 63.88 | 68.05 | 79.45 | 82.55 | 89.15 |
| 0 dB | 47.48 | 52.35 | 60.00 | 66.15 | 75.30 |
| 5 dB | 26.23 | 26.80 | 34.45 | 39.75 | 51.75 |
| 10 dB | 15.6 | 13.95 | 16.30 | 19.75 | 27.70 |
| 15 dB | 9.3 | 7.90 | 7.40 | 9.50 | 9.50 |
| 20 dB | 6.63 | 5.75 | 4.10 | 5.40 | 3.15 |
| Clean | 5.10 | 4.35 | 2.40 | 3.05 | 1.25 |

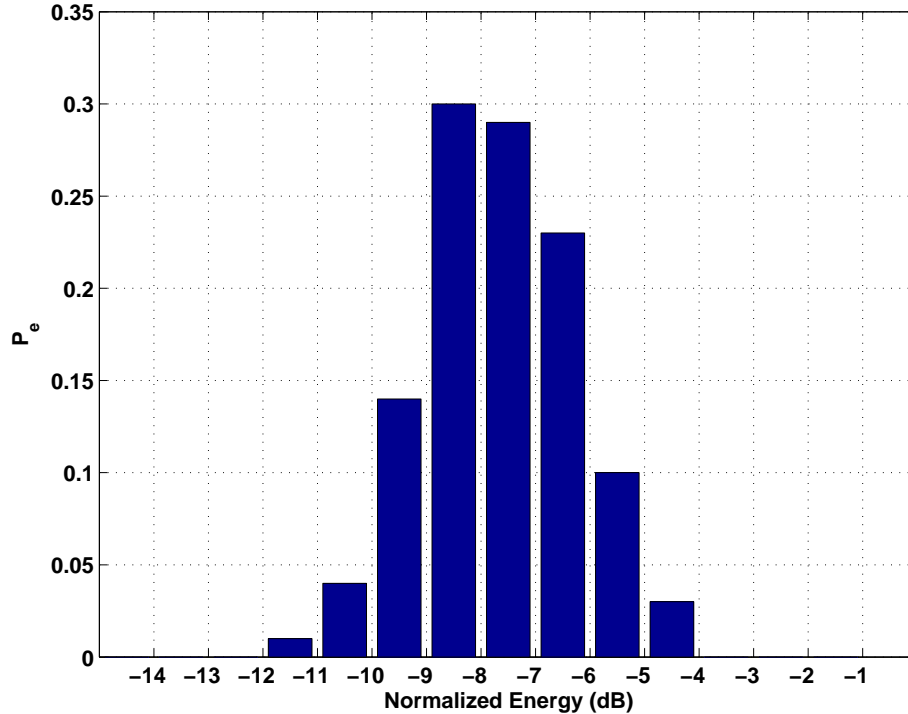
Fig. 24 may be useful for understanding the rationale behind using an energy based voicing detection algorithm. A sample speech utterance from a male speaker is energy normalized as shown in Fig. 24. Clearly, energy rises significantly both in GEMS and speech trajectories when voicing occurs. Therefore, energy rise in the 100-400Hz range is used as the indicator of voicing both for speech and radar signals.

The voiced speech sections are segmented with rectangular boxes in Fig. 24. Although there is a very high match between the voicing decisions from the radar and the speech signals, there are also occasional errors. The accuracy of the GEMS voicing detector is analyzed as follows. The performance of the simulation is measured using ten minutes of speech from clean harvard sentences for which GEMS data is available. Six male and six female speakers are used. Speech energy is normalized and divided into 1 dB intervals for more accurate simulation of the system. Probability of misclassification is calculated for each speech-energy interval by using the voicing feature extracted from the GEMS signal as the ground truth. Misclassification rates are shown in Fig. 25.

7.2.2 Simulating the GEMS-based High-Energy Speech Detection

High-energy speech frames are detected using the clean speech signal using the algorithm described in Section 7.1.6. The speech energy threshold of -5 dB is used as mentioned in Section 7.1.6. The GEMS energy threshold is set to -4 (dB) since it gives the minimum P_e

Figure 23: Probability of error for high-energy speech detection using the GEMS device is shown. Speech energy threshold is set to -5 dB. Possible energy values of normalized speech frames are divided into 1 dB segments, and the error rate is shown separately for each energy interval.



for the -5 dB speech energy threshold as shown in Table 16.

The GEMS signal and the clean speech signal do not always make the same decision in detecting the high energy frames. Misclassification rates are calculated using the method described for the voicing feature. Fig. 23 shows the intervals used for the log-energy of the frames and the probability of making an error for each interval.

7.2.3 Incorporating the Simulation Errors into ASR Experiments

Probability of simulation errors are included in the experimental setup as follows. For each speech frame its corresponding energy interval is found. High-energy and voicing detection are done using the clean speech signal. The detection decision is flipped with a probability of simulation error using the clean speech energy and the P_e data shown in Fig. 23 and 25.

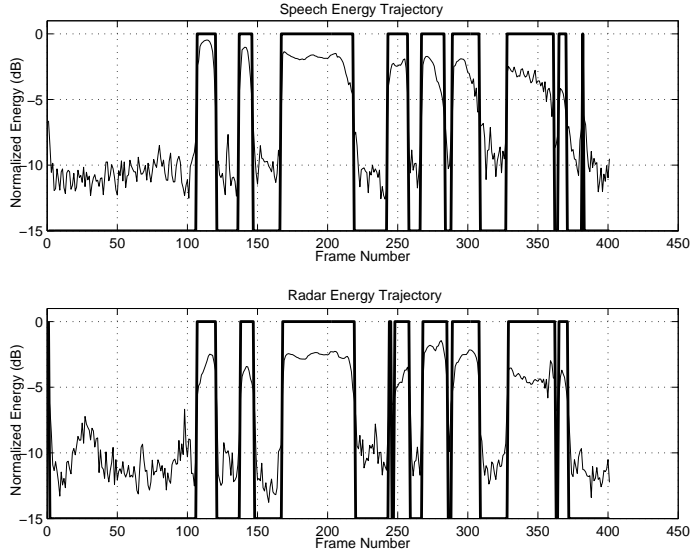


Figure 24: Energy trajectories of the clean speech signal and the radar signal are shown. Energy threshold of -5 dB is used for both cases, and the high-energy segments are shown with rectangular boxes.

Table 16: Probability of error is shown for high-energy speech detection using the GEMS device. Probability of error for various energy thresholds are shown. For each speech energy threshold, the threshold level for the GEMS energy threshold is tuned for the smallest probability of error.

| | | | | |
|--------------------|-------|-------|-------|--------|
| Acoustic Threshold | -5 dB | -7 dB | -9 dB | -11 dB |
| Radar Threshold | -4 dB | -5 dB | -5 dB | -9 dB |
| P_e | 0.09 | 0.05 | 0.12 | 0.15 |

7.3 Experiments

7.3.1 Experimental Setup

The open source ASR software developed at Mississippi State University is used as the baseline recognizer [113]. Words are modeled with 20 state left-to-right Hidden Markov Models (HMM) using 16 Gaussian mixtures per state. The training procedure can be summarized as follows. Baum-Welch training is used with a flat start algorithm for initialization. Single mixture models are trained using Baum-Welch training with four iterations. After the single mixture models are trained, a one state short pause model is trained to represent the short silence segments between words. One mixture models with the short pause model are iterated nine times. Once the single mixture training for all models are completed, the

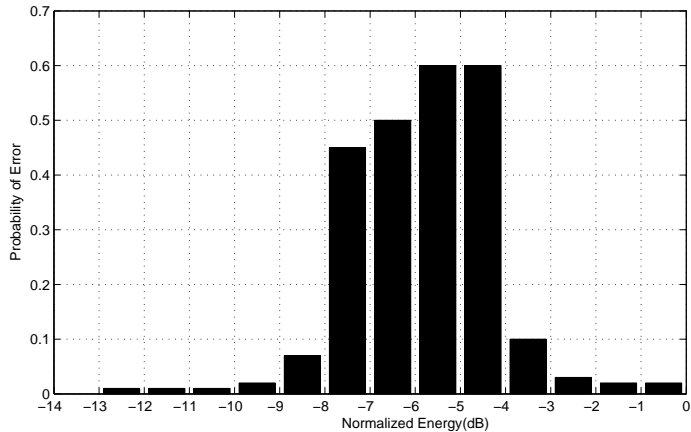


Figure 25: Probability of error is shown for the simulation of voicing detection. Normalized speech energy (dB) is divided into 1 dB intervals, and the error rates are shown for each energy interval.

mixture number is increased progressively by a split strategy where at each step the mixture number is doubled, and new mixtures are retrained by four Baum-Welch iterations.

The Aurora2 database, which contains clean and noisy digits recorded in various noise environments, is used for the experiments [62]. Clean training data is used for training, and the four noisy testing data with subway, babble, car, and exhibition noises are used for testing.

The GEMS data is not yet available for the Aurora2 task. The procedure proposed in Section 7.2 is used for simulating the voicing and high-energy speech detection using the GEMS signal. The errors that occur due to simulation are included both in training and testing.

The baseline system contains 12 MFCC coefficients and energy feature with their first and second order derivatives. Thus, there are 39 features in the baseline system. Energy normalization, described in Section 7.1.6, is performed both for the baseline system and the proposed system.

7.3.2 Results

The performance of the baseline system is compared with the FF and DM methods separately. Table 17 shows the results for the baseline system. Table 18 shows the results with

the DM method, and Table 19 shows the results with the FF method. Improvements over the baseline system using the two methods and the proposed system that uses both methods is shown in Fig. 26.

The DM method provides significant improvement over the baseline system below 15 dB while its performance suddenly and severely drops above 10 dB. The reason for this problem is that at 15 dB even the low-energy speech sounds can be highly reliable, and the DM method cannot take advantage of that reliable information.

The FF method has a different performance gain curve. The gain follows a bell curve with peak at 10 dB. Although the performance gain starts decreasing above 10 dB, it does not experience the sharp drop that is observed with the DM method. However, below 10 dB, its performance gain starts degrading significantly faster than the DM method.

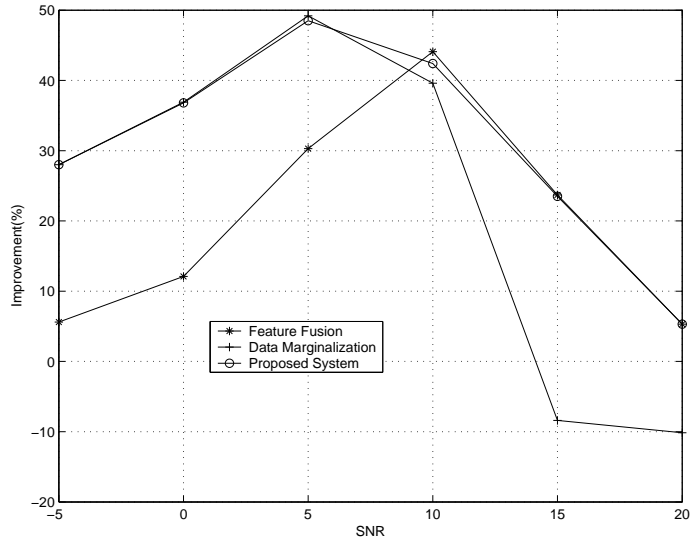


Figure 26: Performance improvement over the baseline system using the feature fusion (FF) method, data marginalization (DM) method, and the proposed system.

The FF and DM methods compliment each other, and the voicing feature is efficiently utilized in the proposed system by using both methods. Hence, the proposed system significantly outperforms the baseline system at all SNRs as shown in Table 20 and Fig. 26.

The proposed system is compared with four other relatively low complexity popular algorithms. Cepstral mean normalization (CMN) is a low complexity feature normalization

Table 17: Performance results in terms of WER for four noise types using the baseline system.

| SNR | Subway | Babble | Car | Exhibition | Avg. |
|-------|--------|--------|------|------------|------|
| -5 dB | 88.6 | 89.7 | 88.2 | 88.2 | 88.7 |
| 0 dB | 75.1 | 75.5 | 76.5 | 73.9 | 75.2 |
| 5 dB | 50.4 | 53.1 | 55.7 | 47.8 | 51.7 |
| 10 dB | 26.0 | 29.4 | 26.8 | 21.4 | 25.9 |
| 15 dB | 9.5 | 9.5 | 7.2 | 8.1 | 8.6 |
| 20 dB | 3.4 | 2.9 | 3.3 | 3.6 | 3.3 |

algorithm that is commonly used in the state-of-the-art systems. It improves the performance over the baseline system at low SNRs, but the improvement is not significant.

An energy-based voice activity detection (VAD) algorithm is used with the baseline system that extracts the voice activity information from the clean speech. Thus, the VAD result in Fig. 27 is the upper bound of its performance. The improvement with VAD is even less than the CMN case. Upper bound of performance is used since the choice of VAD system can significantly affect the performance. It is shown that the gain is small even for the best possible VAD.

The Ephraim-Malah Suppression Rule (EMSR) [41] is used at the front-end, which is a relatively high complexity speech enhancement algorithm. The EMSR system provides a higher performance gain than the other two systems at the expense of increased complexity. However, the proposed system performs significantly better than the EMSR system at all SNRs. Moreover, the EMSR system also has the problem of introducing its own distortion that decreases the performance at high SNRs.

Finally, the proposed system is compared with the marginalization-based missing data system described in [20]. 24 Mel-spaced filters, that are also used in the proposed system, are used to extract the subband features. Spectral subtraction technique is used to detect the missing subbands. Total WER for all four noise types with the subband features are given in Table 22. WER after applying the standard missing data technique is given in Table 21. Clearly, the subband system is significantly intolerant to additive noise compared to the MFCC features. The marginalization method significantly improves the performance

Table 18: Performance results in terms of WER for four noise types using the DM method.

| SNR | Subway | Babble | Car | Exhibition | Avg. |
|-------|--------|--------|------|------------|------|
| -5 dB | 65.6 | 60.3 | 66.4 | 63.2 | 63.9 |
| 0 dB | 48.4 | 41.5 | 52.1 | 47.9 | 47.5 |
| 5 dB | 28.7 | 21.2 | 26.4 | 28.8 | 26.2 |
| 10 dB | 17.3 | 11.0 | 16.8 | 17.5 | 15.6 |
| 15 dB | 9.9 | 7.0 | 10.1 | 10.2 | 9.3 |
| 20 dB | 7.1 | 5.6 | 6.5 | 7.3 | 6.6 |

Table 19: Performance results in terms of WER for four noise types using the data fusion method.

| SNR | Subway | Babble | Car | Exhibition | Avg. |
|-------|--------|--------|------|------------|-------|
| -5 dB | 83.6 | 83.3 | 86.1 | 81.7 | 83.7 |
| 0 dB | 62.9 | 68.6 | 70.6 | 62.6 | 66.2 |
| 5 dB | 34.5 | 37.8 | 38.1 | 33.8 | 36.05 |
| 10 dB | 16.9 | 12.5 | 12.9 | 15.6 | 14.5 |
| 15 dB | 8.3 | 4.1 | 5.0 | 8.8 | 6.6 |
| 20 dB | 3.6 | 2.0 | 2.9 | 4.0 | 3.1 |

in noisy environments although the performance is still far from the proposed technique. The performance drop observed in the proposed system at very high SNRs is also observed for the missing-data system for the clean environment. A significant factor in such low performance is the difficulty in detecting the low SNR bands as mentioned in the first section. The proposed system does not have that problem since the GEMS device can accurately detect the voiced speech segments.

Table 20: Performance results in terms of WER for four noise types using the proposed system.

| SNR | Subway | Babble | Car | Exhibition | Avg. |
|-------|--------|--------|------|------------|------|
| -5 dB | 65.6 | 60.3 | 66.4 | 63.2 | 63.9 |
| 0 dB | 48.5 | 41.6 | 52.1 | 47.9 | 47.5 |
| 5 dB | 29.3 | 21.8 | 26.5 | 29.0 | 26.7 |
| 10 dB | 17.0 | 11.9 | 14.5 | 16.2 | 14.9 |
| 15 dB | 8.3 | 4.1 | 5.0 | 8.8 | 6.6 |
| 20 dB | 3.6 | 2.0 | 2.9 | 4.0 | 3.1 |

Table 21: Performance results in terms of WER for four noise types using the standard missing data system.

| SNR | Subway | Babble | Car | Exhibition |
|-------|--------|--------|------|------------|
| -5 dB | 90.6 | 91.4 | 91.9 | 100 |
| 0 dB | 89.5 | 90.0 | 91.6 | 91.3 |
| 5 dB | 87.9 | 88.2 | 91.5 | 90.6 |
| 10 dB | 78.0 | 77.4 | 82.7 | 83.6 |
| 15 dB | 57.0 | 52.6 | 56.6 | 60.9 |
| 20 dB | 34.5 | 30.4 | 30.4 | 35.3 |
| Clean | 3.1 | 4.3 | 4.5 | 3.3 |

Table 22: Performance results in terms of WER for four noise types using the 24 Mel-spaced filter outputs.

| SNR | Subway | Babble | Car | Exhibition |
|-------|--------|--------|------|------------|
| -5 dB | 100 | 100 | 100 | 100 |
| 0 dB | 100 | 100 | 100 | 100 |
| 5 dB | 100 | 100 | 100 | 100 |
| 10 dB | 98.8 | 90.9 | 99.6 | 99.7 |
| 15 dB | 84.5 | 60.6 | 87.6 | 93.6 |
| 20 dB | 49.9 | 24.3 | 46.3 | 59.6 |
| Clean | 3.1 | 3.1 | 3.3 | 3.3 |

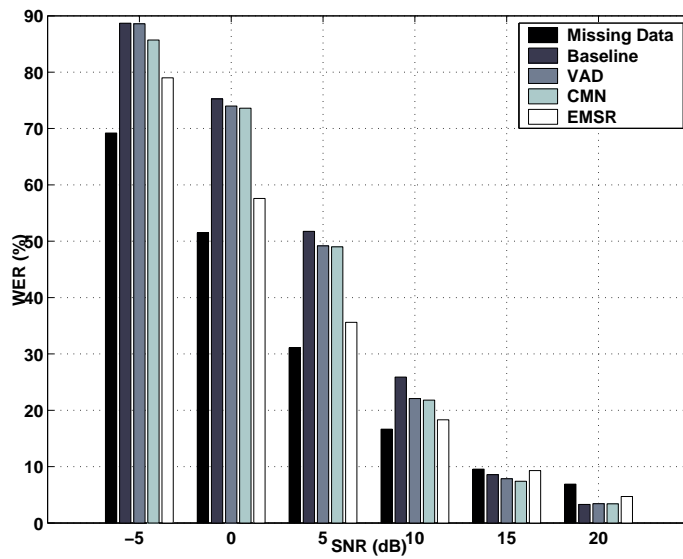


Figure 27: Comparison of the performance between the proposed system, the baseline system, and other relatively low complexity noise-robust systems are shown. The results are averaged over the four noise types of test-a in the Aurora2 database.

CHAPTER VIII

EFFECTS OF SEVERE SENSOR NOISE ON GEMS-BASED FEATURES

Although the GEMS signal is robust to acoustic noise, it typically contains some amount of non-acoustic noise. The noise problem is particularly severe with some of the recordings in the ARCON database for some of the female speakers and some speakers with thicker neck skin.

The noise-robust ASR methods presented in the previous chapter were tested using a simulated GEMS feature derived from clean speech. However, this method ignores the severe non-acoustic noise that existed in some of the GEMS data since the GEMS recordings with severe non-acoustic noise were not used in the process of determining the simulation parameters.

In this chapter, effects of noise on the GEMS-based features are discussed with a focus on the severely noisy cases. Effects of noise on the three features, periodicity, energy, and harmonicity, are discussed in the next section. In Section 8.2, a signal enhancement technique and a missing data technique are investigated to alleviate the effects of sensor noise on the feature fusion based multisensor ASR. Experimental results are presented in Section 8.3.

8.1 Noise Interference in the GEMS Signal

Noise in the GEMS signal is modelled with additive white noise. Hence, the noisy GEMS signal is

$$g_n(i) = g(i) + n(i) \tag{60}$$

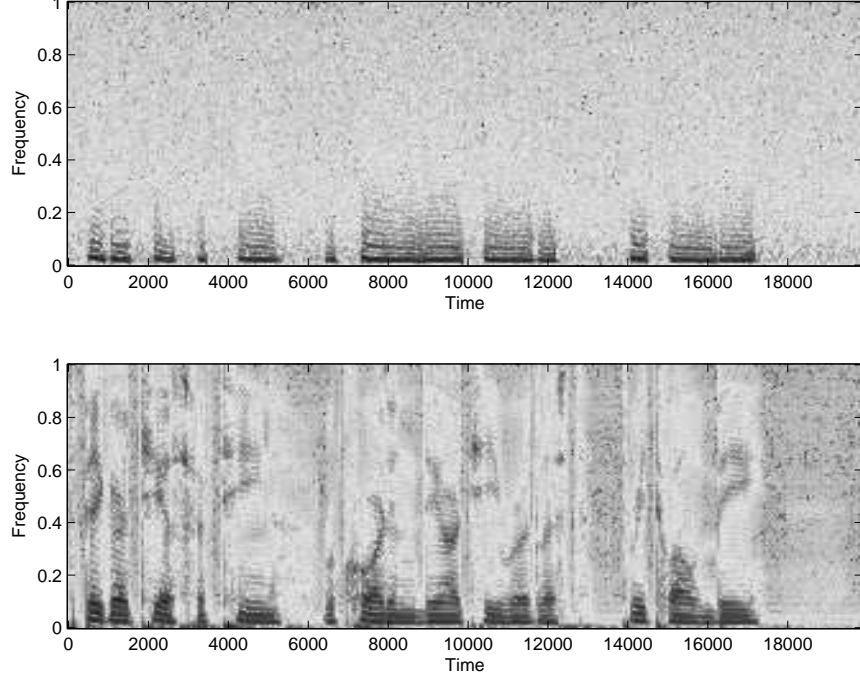


Figure 28: Spectrogram of a clean speech signal is shown in the top figure. Spectrogram of the corresponding noisy GEMS signal is shown in the bottom Figure.

where $g(i)$ is the clean GEMS signal, and $n(i)$ is the noise signal at the time sample i . The magnitude spectrum of the noisy GEMS signal is

$$G_n(k) = G(k) + N(k) \quad (61)$$

at the frequency sample k .

Noise is typically stationary for a given recording in the ARCON database. In Fig. 28, a sample noisy GEMS spectrum is shown with the corresponding audio for a sample utterance. Effects of noise on the three features derived from the GEMS signal, periodicity, energy, and harmonicity features defined in Section 2.4, are discussed below. Short-time analysis of the GEMS signal is used in the discussions.

Trajectory of the periodicity feature extracted from a sample noisy GEMS signal is shown in Fig. 29. Periodicity of a noisy GEMS signal at time frame j is

$$v_{n,j} = \arg \max_{2.5 \text{ msec} < \tau_j < 10 \text{ msec}} R_{gg}(\tau_j) + R_{nn}(\tau_j) \quad (62)$$

where $R_{nn}(\tau_j)$ is the autocorrelation function of the additive noise, and $R_{gg}(\tau_j)$ is the

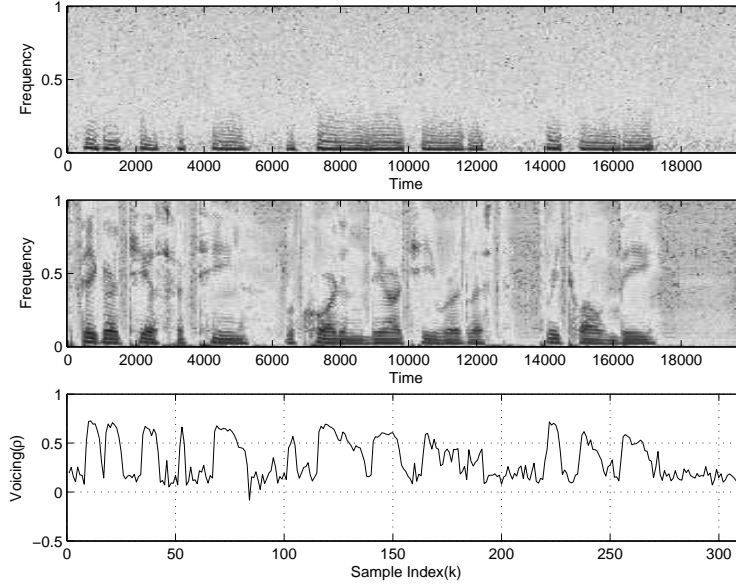


Figure 29: Spectrogram of a sample GEMS signal is shown in the top figure. Corresponding speech spectrogram is shown in the middle figure. Trajectory of the voicing features extracted from the GEMS signal is shown in the bottom figure. Speech and GEMS samples are taken from the DRT sequences in the ARCON database recorded in the quiet conditions by a female speaker.

autocorrelation function of the GEMS signal in the j^{th} frame. Since the noise is white,

$$E\{v_{n,j}\} = \arg \max_{2.5 \text{ msec} < \tau_j < 10 \text{ msec}} R_{gg}(\tau_j) \quad (63)$$

where $E\{v_{n,i}\}$ is the expected value of the periodicity feature. Thus, on average, the periodicity feature is robust to additive sensor noise.

GEMS signal has a harmonic spectrum similar to the voiced speech spectrum as discussed in Section 2.4. Harmonicity at high frequencies is more sensitive to the sensor noise compared to the low frequency harmonics as shown in Fig. 31 since the noise is white and the high-frequency harmonics have significantly lower signal energy compared to the low-frequency harmonics.

Trajectory of the energy features extracted from a noisy GEMS signal sample is shown in Fig. 30. Similar to the periodicity feature, energy feature is robust to sensor noise if the GEMS signal is not severely contaminated by additive noise. The relative robustness of the energy feature can be explained as follows. The GEMS spectrum has a rapidly decaying spectral envelope as shown in Fig. 31. Although noise can mask a significant

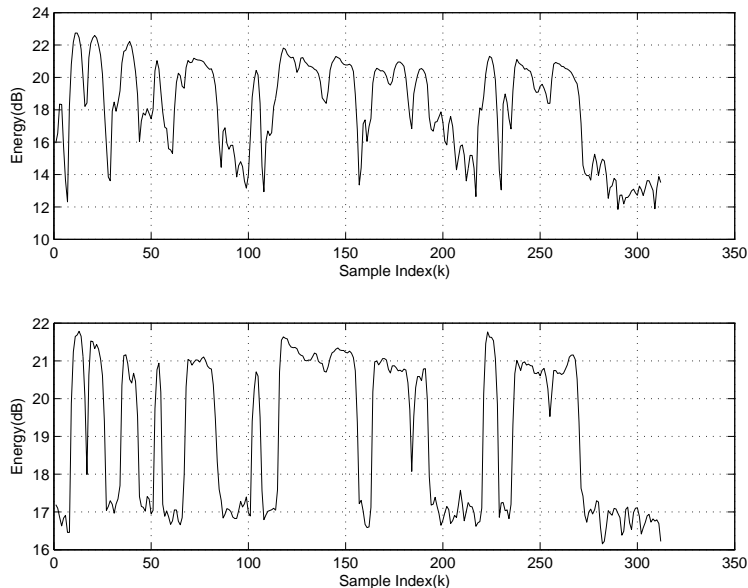


Figure 30: Energy trajectory of a sample speech utterance is shown in Figure a and energy trajectory of the corresponding GEMS signal is shown in Figure b. Speech and GEMS samples are taken from the DRT sequences in the ARCON database recorded in the quiet conditions by a female speaker.

portion of the higher frequencies, the first couple of harmonics, especially the first harmonic at the fundamental frequency, still have significantly greater energy than the sensor noise in general. Therefore, if the GEMS signal is not severely noisy, total energy of the GEMS harmonics masks the total noise energy in a given GEMS signal frame.

8.2 *Feature-Fusion Based ASR with Noisy GEMS Signal*

Effects of the sensor noise on the ASR performance is measured using the word error rate (WER) metric. Two types of features, the energy and periodicity features defined in Section 2.4, are extracted from the GEMS signal. The GEMS-based feature is augmented with the speech-based MFCC features using the feature-fusion method described in Chapter 7.

Two approaches are used to alleviate the effects of noise on the ASR performance. In the first approach, the MMSE estimator proposed in [40] is used to clean the noise in the GEMS signal, and ASR results are reported for the features extracted from the enhanced GEMS (e-GEMS) signal. Both training and testing data are enhanced in this approach.

The GEMS signal is known to be severely noisy for some of the utterances as discussed

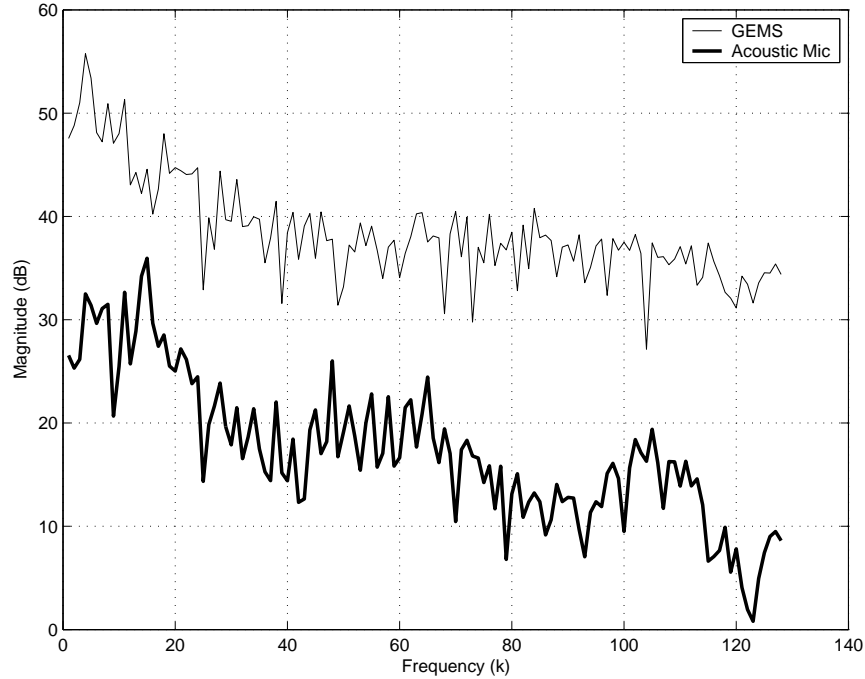


Figure 31: Magnitude spectrums of a noisy GEMS signal and acoustic signal are compared for a voiced speech frame.

in the previous section. In the second approach, a missing-data method is used to minimize the adverse effects of those severely noisy case. The GEMS-based feature is labeled as “missing” if the GEMS signal is severely noisy. The marginalization-based missing data technique discussed in Chapter 7 and [20] is used.

8.3 Experiments

The ARCON database, described in Section 2.4, is reorganized for speech recognition experiments. A corpus that contains speaker-dependent, isolated monosyllables is prepared where each utterance contains a single monosyllable. Ten female and nine male speakers are enrolled in the experiments. Each speaker has seven minutes of training data and three minutes of testing data. Hence, a total of 133 minutes of training data and 57 minutes of testing data is collected from all 19 speakers. Speaker M8 in the ARCON database is excluded because he is found to have a voice disorder. The ARCON database, described in Section 2.4, is used for experiments. The DRT data in the ARCON database is reorganized

for speech recognition experiments as described below.

The lexicon is prepared as follows. Pronunciations of all the monosyllables are gathered from the pronunciation dictionary prepared at Carnegie Melon University. All monosyllables are found to be correct except "MIT" which is pronounced as /m/-/iy/-/t/ in the DRT list but as /em/-/ay/-/t/-/iy/ in the dictionary. The final lexicon contains 228 words.

Both speech and GEMS signals are downsampled to 8 kHz, and a 256 sample Hanning window with a frame rate of 128 samples/frame is used to extract the features synchronously from the speech and GEMS samples. Half of the training and testing data for each user is recorded in quiet conditions while the other half is recorded in an office environment that has high SNR. Both acoustic data and GEMS data are prepared for each utterance. GEMS and acoustic signals are synchronized using the method described in Section 2.4.3.

Both speech and GEMS signals are downsampled to 8 kHz, and a 256 sample Hanning window with a frame rate of 128 samples/frame is used to extract the features synchronously from the speech and GEMS samples.

Experiments are performed using the HMM-based Automatic Speech Recognition (ASR) software developed at Mississippi State University. 16 Gaussians per state are used for Gaussian mixture modelling. Syllable templates are used to model each monosyllable. HMM-models for each monosyllable contain 10 states.

8.3.1 Results and Discussion

The baseline system uses the original 39 dimensional MFCC features. The proposed system uses an additional feature, periodicity or energy, extracted from the GEMS signal. Several experiments are performed to compare the performance of the baseline system and the proposed system. The first set of experiments are done to compare the performance of the direct feature fusion method with the baseline system. The results are illustrated in Table 23. Augmenting the GEMS feature with the MFCC features has an adverse effect on the accuracy for both the periodicity and the energy features. The periodicity feature outperforms the energy feature.

The second set of experiments are done by enhancing the GEMS signal with the MMSE

estimator. Both training and testing data are enhanced using the MMSE estimator. Results are shown in Table 24. After signal enhancement, accuracy is approximately nine percent lower in both cases relative to the baseline case. Moreover, accuracy using the e-GEMS signal is lower compared to the accuracy using the GEMS signal.

In the third set of experiments, a missing data technique is used that discards a GEMS utterance if the estimated average noise energy in the utterance is above a hard threshold. The algorithm works as follows. Energy features that are extracted from a GEMS utterance are normalized with a constant multiplication factor k . k is chosen such that the highest energy frame in an utterance has 0 dB energy. The first 10 frames in a given GEMS signal are assumed to be noise-only, and average noise energy is estimated by averaging the energy of the first 10 frames. If the noise energy is found to be above an energy threshold, then the signal is classified as severely noisy.

The energy threshold is determined by maximizing the rejection ratio of the severely noisy GEMS signals while minimizing the loss of other utterances. Severely noisy utterances are hand-selected by inspection of the spectrograms of the utterances in the DRT database. An utterance is labeled as severely noise when noise masks most of the harmonics below 500 Hz. A severely noisy utterance from the ARCON database is shown in Fig. 32.

Pdf of the noise energy estimated in all utterances, and the pdf of the noise energy estimated in the severely noisy utterances are shown in Fig. 33. Analyzing the pdf curves, energy threshold of -10 dB is chosen as the operating point. Therefore, if the noise energy is above -10 dB, then the GEMS features are labeled as missing. If the noise energy is above -10 dB, then the GEMS features are augmented with the MFCC features.

Results with the missing data technique are shown in Table 25. Accuracy of the system is higher than the baseline for both features. The relative improvement is approximately six percent for the energy feature and nine percent for the periodicity feature compared to the baseline case. Thus, these results suggest that the severely noisy GEMS features can significantly degrade the performance of an ASR system when fused with the MFCC features. When those severely noisy features are discarded using a missing data method, GEMS-based features can help increase the accuracy of the clean MFCC features.

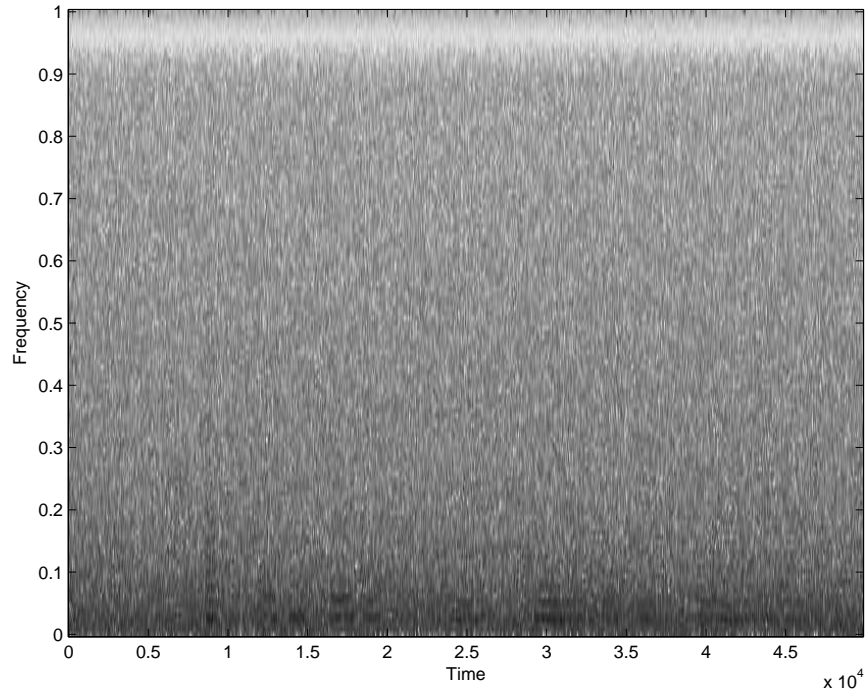


Figure 32: Spectrogram of a severely noisy GEMS sample.

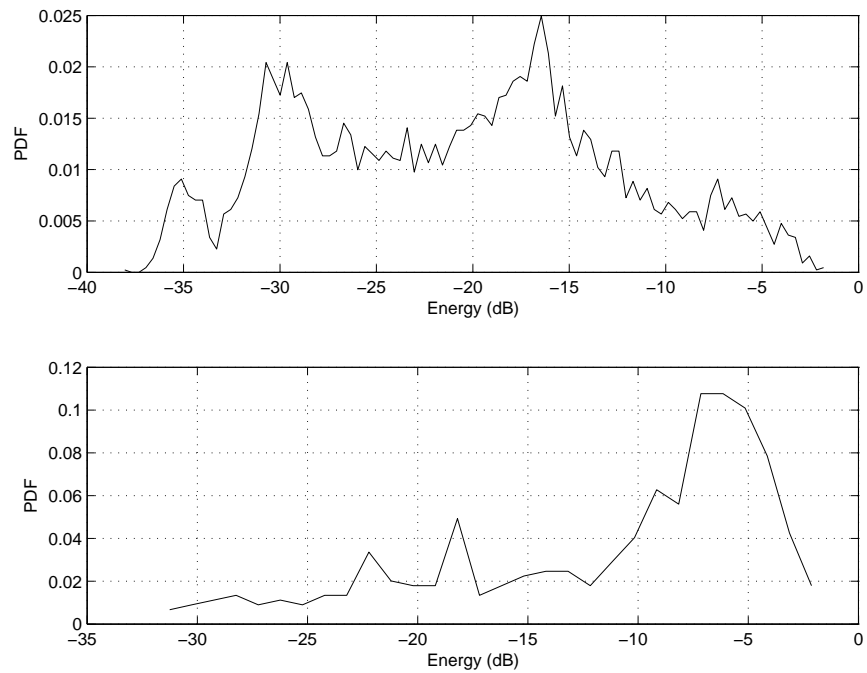


Figure 33: Pdf of the energy features extracted from all (noisy and clean) GEMS utterances are shown in the top figure. Pdf of the energy features extracted from the severely noisy utterances are shown in the bottom figure.

Table 23: A GEMS-based feature is fused with the MFCC features. Two types of features, periodicity and energy, are extracted from the GEMS signal. Results for both features are shown and compared with the MFCC features. Error rates for three different error types, substitution, deletion, insertion, and their total are shown.

| Error Type | MFCC | MFCC and GEMS-energy | MFCC and GEMS-periodicity |
|--------------|------|----------------------|---------------------------|
| Substitution | 11.4 | 11.6 | 11.0 |
| Deletion | 0.2 | 0.3 | 0.3 |
| Insertion | 0.0 | 0.0 | 0.0 |
| Total Error | 11.6 | 11.9 | 11.7 |

Table 24: ASR performance when the GEMS signal is enhanced with the MMSE estimator. Error rates for three different error types, substitution, deletion, insertion, and their total are shown. GEMS-based feature is fused with the MFCC feature vector. Two types of features, periodicity and energy, are extracted from the GEMS signal. Results for both features are shown.

| Error Type | MFCC and e-GEMS-energy | MFCC and e-GEMS-periodicity |
|--------------|------------------------|-----------------------------|
| Substitution | 12.0 | 11.8 |
| Deletion | 0.2 | 0.2 |
| Insertion | 0.0 | 0.0 |
| Total Error | 12.2 | 12.0 |

Table 25: ASR performance when the missing data technique is used for the GEMS-based features. Error rates for three different error types, substitution, deletion, insertion, and their total are shown. GEMS-based feature is fused with the MFCC feature vector. Two types of features, periodicity and energy, are extracted from the GEMS signal. Results for both features are shown.

| Error Type | MFCC and MD-GEMS-energy | MFCC and MD-GEMS-periodicity |
|--------------|-------------------------|------------------------------|
| Substitution | 10.6 | 10.0 |
| Deletion | 0.3 | 0.3 |
| Insertion | 0.0 | 0.0 |
| Total Error | 10.9 | 10.3 |

CHAPTER IX

CONCLUSION

This thesis presents novel multi-sensor speech enhancement methods to improve the intelligibility of MELP-encoded speech. A perceptually-motivated Bayesian risk minimization (BRM) framework is proposed for speech enhancement. The proposed framework takes into account the sound-class dependence of the human auditory perception. Misestimations due to estimation uncertainty and SNR uncertainty are analyzed using the robust Bayesian analysis (RBA). Heuristically-motivated spectral subtraction based systems are analyzed under the proposed theoretical framework.

Using the ideas presented in the RBM framework, a segmentation-based, multi-sensor speech enhancement system is proposed that is shown to improve the intelligibility of encoded speech in M2 tank noise environment. The proposed system is tested by using it in tandem with a 2400 bps MELP speech coder, and comparing with a state-of-the-art 2400 bps MELPe coder using the DRT intelligibility test. Concatenation of the MELP coder and the proposed system is shown to have higher intelligibility for nasality, sibilant, and voicing features compared to the MELPe coder.

The segmentation-based system uses a novel aharmonic comb filter (ICF) in addition to the acoustic-phonetic knowledge provided by a segmentation algorithm. The ICF system uses the GEMS signal to detect perceptually important high-signal-power (HSP) frequencies in the voiced speech spectrum. HSP locations are enhanced with the MMSE estimator while the non-HSP locations are severely suppressed. Improvement in the perceptual quality compared to the MMSE system used in MELPe is shown using A/B subjective tests. Moreover, the ICF system also improved the segmental-SNR compared to the MMSE system.

The ICF system creates a sparse spectrum by suppressing the low-energy non-HSP frequencies. LPC extraction from such a sparse spectrum is analyzed in detail in Chapter 5. It is shown that suppressing the non-HSP frequencies without any further enhancement of

Table 26: The quality and intelligibility tests that have been done to measure the performance of the proposed enhancement systems.

| | Quality Test | Intelligibility Test |
|---------------------------|--------------|----------------------|
| ACF system | A/B test | None |
| Segmentation-based system | None | DRT |

the HSP frequencies can improve the LPC spectrum significantly.

A multi-sensor segmentation system developed at Georgia Tech is used for segmentation in the experiments. Multi-sensor, HMM-based speech segmentation/classification is also considered. Because of the lack of GEMS data, in the first phase, a voicing feature, extracted from clean speech, is used for the Aurora2 speech recognition task. The voicing feature is exploited by feature fusion and data marginalization methods. Feature fusion method is found to be more effective at higher SNRs while the data marginalization method is found to be effective at lower SNRs. Therefore, a combined system is proposed that uses the data marginalization method at low SNRs and the feature fusion method at high SNRs. The combined system is shown achieve significant gain over the baseline system and the individual systems. Moreover, the proposed system outperformed some of the popular noise-robust methods. The tests that have been done to measure the effectiveness of the proposed algorithms are shown in Table 26.

In the second phase of the noise-robust segmentation research, an isolated monosyllable database is created using the ARCON DRT database. Feature-fusion method is used where the GEMS-derived feature is fused with the MFCC features. Energy and voicing features are used. Experiments show an increase in word error rate compared to the baseline system when the GEMS-based feature is fused with the MFCC features. An investigation of the reason showed that the GEMS signal can be severely noisy in some situations because of misplacement of the sensor, amount of skin fat, or gender differences. These severely noisy cases are found to degrade the performance. Two methods are used to alleviate the effects of severely noisy GEMS data. In the first method, signal enhancement using an MMSE estimator is considered. However, signal enhancement degraded the accuracy even further. In the second method, a data marginalization based missing data method is used where

severely noisy GEMS utterances are labeled as missing data. The data marginalization method is found to improve the accuracy of the system over the baseline system. Periodicity feature is found to perform better than the energy feature in all experiments.

REFERENCES

- [1] AGARWAL, T. and KABAL, P., "Pre-processing of noisy speech for voice coders," in *IEEE Workshop on Speech Coding*, (Ibaraki, JAPAN), Oct 2002.
- [2] AKBACAK, M. and HANSEN, J. H. L., "Environmental sniffing: noise knowledge estimation for robust speech systems," in *ICASSP*, (Hong Kong), Apr 2003.
- [3] ANDERSON, D. V. and CLEMENTS, M. A., "Audio signal noise reduction using multi-resolution sinusoidal modeling," in *ICASSP*, (Phoenix, Arizona), Mar 1999.
- [4] *ANSI S3.2-1989*, ch. Method for Measuring the Intelligibility of Speech Over Communication Systems. New York: American Standards Association.
- [5] ARSLAN, L., MCCREE, A., and VISWANATHAN, V., "New methods for adaptive noise suppression," in *ICASSP*, (Detroit, Michigan), May 1995.
- [6] BAHOURA, M. and ROUAT, J., "Wavelet speech enhancement based on the teager energy operator," *IEEE Transactions on Signal Processing*, vol. 8, Jan 2001.
- [7] BARNWELL, T., CLEMENTS, M. A., ANDERSON, D. V., MOORE, E., LEE, M., ERTAN, A. E., KRISHNAN, V., KAMATH, S., CHOI, W., HU, J., DEMIROGLU, C., WHITEHEAD, P. S., and DUREY, A. S., "Low bit rate coding of speech in harsh conditions using non-acoustic auxiliary devices," in *Special Workshop in Maui: Lectures by masters in Speech Processing*, (Maui, Hawaii), Jan 2004.
- [8] BURNETT, G. C., *The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract*. PhD thesis, University of California Davis, 1999.
- [9] CAPPE, O., "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, Apr 1994.
- [10] CERISARA, C., "Towards missing data recognition with cepstral features," in *EUROSPEECH*, (Geneva, Switzerland), Sep 2003.
- [11] CHEN, J., PALIWAL, K. K., and NAKAMURA, S., "Cepstrum derived from differentiated power spectrum for robust speech recognition," *Speech Communication*, vol. 41, 2003.
- [12] CHENG, Y. M. and O'SHAUGHNESSY, D., "Speech enhancement based conceptually on auditory evidence," *IEEE Transactions on Signal Processing*, vol. 39, Sep 1991.
- [13] CHILTON, E. and MARVI, H., "Two-dimensional root cepstrum as feature extraction method for speech recognition," *Electronics Letters*, vol. 39, May 2003.

- [14] COHEN, I., “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, Sep 2001.
- [15] COHEN, I., “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 9, 2002.
- [16] COHEN, I., “On the decision-directed estimation approach of Ephraim and Malah,” in *ICASSP*, (Quebec, CA), May 2004.
- [17] COHEN, I., “Speech enhancement using a noncausal a priori SNR estimator,” *IEEE Signal Processing Letters*, vol. 11, Sep 2004.
- [18] COLLURA, J. S., “A duration-based confidence measure for automatic segmentation of noise corrupted speech,” in *ICSLP*, (Sydney, Australia), Dec 1998.
- [19] COLLURA, J. S., “A segment-based algorithm of speech enhancement for robust speech recognition,” in *EUROSPEECH*, (Geneva, Switzerland), Oct 2003.
- [20] COOKE, M. P., GREEN, P. D., JOSIFOVSKI, L., and VIZINHO, A., “Robust automatic speech recognition with missing and unreliable acoustic data,” *Speech Communication*, vol. 34, 2001.
- [21] CUI, X., ISELI, M., ZHU, Q., and ALWAN, A., “Evaluation of noise robust features on the Aurora databases,” in *ICSLP*, (Denver, Colorado), Sep 2002.
- [22] DAVILA, C. E., “A subspace approach to estimation of autoregressive parameters from noisy measurements,” *IEEE Transactions on Signal Processing*, vol. 46, Feb 1998.
- [23] DAVILA, C. E., “Speech enhancement using linear prediction residual,” *Speech Communication*, vol. 28, 1999.
- [24] DEISHER, M. and SPANIAS, A., “HMM-based speech enhancement using harmonic modeling,” in *ICASSP*, (Munich, Germany), Apr 1997.
- [25] DELATTRE, P. C. and COOPER, F. S., “Acoustic loci and transitional cues for consonants,” *The Journal of the Acoustical Society of America*, vol. 27, Jul 1955.
- [26] DELATTRE, P. C. and COOPER, F. S., “Some cues for the distinction between voiced and voiceless stops in initial position,” *Language and Speech*, vol. 1, pp. 153–167, Jul-Sep 1958.
- [27] DELLER, J., PROAKIS, J., and HANSEN, J., *Discrete-Time Processing of Speech Signals*. New York:Macmillan: Wiley-IEEE Press, 1993.
- [28] DEMIROGLU, C. and ANDERSON, D. V., “A soft decision MMSE amplitude estimator as a noise preprocessor to speech coders using a glottal sensor,” in *ICSLP*, (Jeju island, Korea), Oct 2004.
- [29] DEMIROGLU, C. and BARNWELL, T. P., “A missing-data approach to noise-robust LPC extraction for voiced speech using auxiliary sensors,” in *ICASSP*, vol. 1, (Philadelphia, PA), pp. 773 – 776, Mar 2005.

- [30] DEMIROGLU, C., KAMATH, S., and ANDERSON, D. V., “Segmentation-based noise suppression using auxiliary sensors,” in *ASILOMAR*, (Pacific Grove, CA), Nov 2004.
- [31] DEMUYNCK, K. and LAUREYS, T., “A comparison of different approaches to automatic speech segmentation,” in *In Proc. 5th International Conference on Text, Speech and Dialogue*, (Brno, Czech Republic), Sep 2002.
- [32] DENG, L., DROPPA, J., and ACERO, A., “Exploiting variances in robust feature extraction based on a parametric model of speech distortion,” in *ICSLP*, (Denver, Colorado), Sep 2002.
- [33] DERICHE, M., “AR parameter estimation from noisy data using the EM algorithm,” in *ICASSP*, (Adelaide, Australia), Apr 1994.
- [34] DRUCKER, H., “Speech processing in a high ambient noise environment,” *IEEE Trans. Audio Electroacoust.*, vol. AU-16, Jun 1968.
- [35] EPHRAIM, Y., “On minimum mean square error speech enhancement,” in *ICASSP*, (Albuquerque), Apr 1990.
- [36] EPHRAIM, Y., “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Transactions on Signal Processing*, vol. 40, Apr 1992.
- [37] EPHRAIM, Y., “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, Oct 1992.
- [38] EPHRAIM, Y. and COHEN, I., *Recent advancements in speech enhancement*. The Electronic Handbook, CRC Press, 2005.
- [39] EPHRAIM, Y., LEV-ARI, H., and ROBERTS, W. J. J., *A Brief Survey of Speech Enhancement*. The Electrical Engineering Handbook, CRC Press to appear, 2005.
- [40] EPHRAIM, Y. and MALAH, D., “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and SignalProcessing*, vol. ASSP-32, Dec 1984.
- [41] EPHRAIM, Y. and MALAH, D., “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and SignalProcessing*, vol. ASSP-33, 1985.
- [42] EPHRAIM, Y., MALAH, D., and JUANG, B.-H., “On the application of hidden Markov models for enhancing noisy speech,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, Dec 1989.
- [43] EPHRAIM, Y. and TREES, H. L. V., “A signal subspace approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, Jul 1995.
- [44] ERTAN, A. E., *Pitch Synchronous Analysis of Speech Signal for Improving the Quality of Low Bit Rate Speech Coders*. PhD thesis, Georgia Institute of Technology, 2003.
- [45] EVANS, N. W. D. and MASON, J. S., “Computationally efficient noise compensation for robust automatic speech recognition assessed under the Aurora 2/3 framework,” in *ICSLP*, (Denver, CO), Sep 2002.

- [46] FEIJOO, S. and FERNANDEZ, S., “Location of spectral cues for the spectral identification of fricatives,” in *Forum Acusticum*, (Spain), Sep 2002.
- [47] FRAZIER, R. H., SAMMATT, S., BRAIDA, L. D., and OPPENHEIM, A. V., “Enhancement of speech by adaptive filtering,” in *ICASSP*, (Philadelphia), 1976.
- [48] GALES, M. J. F., “Robust continuous speech recognition using parallel model combination,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4, 1996.
- [49] GALES, M. J. F., “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, 1998.
- [50] GALES, M. J. F. and WOODLAND, P. C., “Mean and variance adaptation within the MLLR framework,” *Computer Speech and Language*, vol. 10, 1996.
- [51] GELBART, D. and MORGAN, N., “Evaluating long-term spectral subtraction for reverberant ASR,” in *ASRU*, (Trento, Italy), 2001.
- [52] GONG, Y., “Speech recognition in noisy environments: A survey,” *Speech Communication*, vol. 16, 1995.
- [53] GRACIARENA, M., FRANCO, H., SONMEZ, K., and BRATT, H., “Combining standard and throat microphones for robust speech recognition,” *IEEE Signal Processing Letters*, vol. 10, 2003.
- [54] GRAY, A. H. and MARKEL, J. D., “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, Oct 1976.
- [55] GUILMIN, G., JEANNES, R. L. B., and GOURNAY, P., “On the optimization of noise pre-processing for a low bit rate parametric speech coder,” in *COST 254 workshop on friendly exchanging through the Net*, (Bordeaux), Mar 2000.
- [56] HAMME, H. V., “Two correction models for likelihoods in robust speech recognition using missing feature theory,” in *EUROSPEECH*, (Geneva, Switzerland), Sep 2003.
- [57] HANSEN, J. H. L., SARIKAYA, R., YAPANEL, U., and PELLOM, B., “Robust digit recognition in noise: An evaluation using the SPINE corpus,” in *EUROSPEECH*, (Aalborg, Denmark), Sep 2001.
- [58] HARRIS, K. S., HOFFMAN, H. S., and GRIFFITH, B. C., “The discrimination of speech sounds within and across phoneme boundaries,” *Journal of Experimental Psychology*, vol. 54, no. 5, pp. 358–368, 1958.
- [59] HASAN, M. K., FATTAH, S. A., and KHAN, R., “Identification of noisy ar systems using damped sinusoidal model of autocorrelation function,” *IEEE Signal Processing Letters*, vol. 10, Jun 2003.
- [60] HERMAN, H. and MORGAN, N., “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, Oct 1994.
- [61] HERNANDO, J. and NADEU, C., “Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, Jan 1997.

- [62] HIRSCH, H. G. and PEARCE, D., “The Aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium*, (Paris, France), Sep 2000.
- [63] HOLMES, J. N. and SEDGWICK, N., “Noise compensation for speech recognition using probabilistic models,” in *Proceedings of ICASSP*, (Tokyo), Apr 1986.
- [64] HU, H., KUO, F., and WANG, H., “Supplementary schemes to spectral subtraction for speech enhancement,” *Speech Communication*, vol. 36, 2002.
- [65] HU, R. and ANDERSON, D. V., “Single acoustic-channel speech enhancement based on glottal correlation using non-acoustic sensor,” in *ICSLP*, (Jeju island, Korea), Oct 2004.
- [66] HU, Y. and LOIZOU, P. C., “A generalized subspace approach for enhancing speech corrupted by colored noise,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, Jul 2003.
- [67] HUNT, M. J., “Spectral signal processing for ASR,” in *ASRU-99*, (Keystone, Colorado), Dec 1999.
- [68] JABLOUN, F. and CHAMPAGNE, B., “Incorporating the human hearing properties in the signal subspace approach for speech enhancement,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, Nov 2003.
- [69] JENSEN, J. and HANSEN, J. H. L., “Speech enhancement using a constrained iterative sinusoidal model,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, Oct 2001.
- [70] JUNEJA, A. and ESPY-WILSON, C., “Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines,” in *Proceedings of International Joint Conference on Neural Networks*, (Portland, Oregon), Jul 2003.
- [71] KAMATH, S. and LOIZOU, P., “Multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *ICASSP*, (Orlando, FL), May 2002.
- [72] KAY, S. M., “All-pole modeling of degraded speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-26, Jun 1978.
- [73] KAY, S. M., “The effect of noise on the autoregressive spectral estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, Oct 1979.
- [74] KAY, S. M., “Noise compensation of autoregressive spectral estimates,” *IEEE Transactions on Speech and Audio Processing*, vol. 28, pp. 292–303, 1980.
- [75] KIM, N. S. and CHANG, J.-H., “Spectral enhancement based on global soft decision,” *IEEE Signal Processing Letters*, vol. 7, May 2000.
- [76] KIM, Y. and CONKIE, A., “Automatic segmentation combining an HMM-based approach and spectral boundary correction,” in *ICSLP*, (Denver, Colorado), Sep 2002.
- [77] KLATT, D. H., “A digital filter bank for spectral matching,” in *Proceedings of ICASSP*, (Washington DC), Apr 1979.

- [78] KLEIN, M. and KABAL, P., “Signal subspace speech enhancement with perceptual post-filtering,” in *ICASSP*, (Orlando, FL), May 2002.
- [79] LAVINE, M., “Sensitivity in Bayesian statistics: the prior and the likelihood,” *Journal of the American Statistical Association*, vol. 86, 1991.
- [80] LEE, C. H., “On stochastic feature and model compensation approaches to robust speech recognition,” *Speech Communication*, vol. 25, 1998.
- [81] LEE, J. J., LEE, J. H., and LEE, K. Y., “Efficient speech enhancement based on left-right HMM with state sequence detection using LRT,” in *EUROSPEECH*, (Geneva, Switzerland), Oct 2003.
- [82] LEV-ARI, H. and EPHRAIM, Y., “Extension of the signal subspace speech enhancement approach to colored noise,” *IEEE Signal Processing Letters*, vol. 10, Apr 2003.
- [83] LIM, J. S. and OPPENHEIM, A. V., “Enhancement and bandwidth compression of noisy speech,” *Proceedings of IEEE*, vol. 67, Dec 1979.
- [84] LIM, J. S., OPPENHEIM, A. V., and BRAIDA, L. D. V., “Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition,” *IEEE Transactions on acoustics, speech, and signal processing*, vol. ASSP-26, Aug 1978.
- [85] LIN, C., NEIN, H., and HWU, J., “Ga-based noisy speech recognition using two-dimensional cepstrum,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, Nov 2000.
- [86] LJOLJE, A., “Speech recognition using fundamental frequency and voicing in acoustic modeling,” in *ICSLP*, (Denver, CO), Sep 2002.
- [87] LOCKWOOD, P. and BOUDY, J., “Experiments with a nonlinear spectral subtractor (NSS) hidden Markov models and the projection, for robust speech recognition in cars,” *Speech Communication*, vol. 11, 1992.
- [88] LOGAN, B. T. and ROBINSON, A. J., “Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal,” in *ICSLP*, (Yokohama, Japan), Sep 1994.
- [89] MALAH, D., COX, R. V., and ACCARDI, A. J., “Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments,” in *ICASSP*, (Phoenix, AZ), Mar 1999.
- [90] MARTIN, R., “Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors,” in *ICASSP*, (Orlando, FL), May 2002.
- [91] MARTIN, R. and COX, R. V., “New speech enhancement techniques for low bit rate speech coding,” in *IEEE Speech Coding Workshop*, Jun 1999.
- [92] MARTIN, R., MALAH, D., COX, R., and ACCARDI, A., “A noise reduction preprocessor for mobile voice communication,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1046 – 58, Jul 2004.
- [93] MARTIN, R., WITTKE, I., and JAX, P., “Optimized estimation of spectral parameters for the coding of noisy speech,” in *ICASSP*, (Istanbul, Turkey), Jun 2000.

- [94] MASHAO, D. J., GOTOH, Y., and SILVERMAN, H. F., “Analysis of lpc/dft features for an HMM-based alphadigit recognizer,” *IEEE Signal Processing Letters*, vol. 3, Apr 1996.
- [95] MATROUF, D. and GAUVIN, J. L., “Using an HMM state-dependent filtering for speech enhancement,” in *ICASSP*, (Phoenix, Arizona), Mar 1999.
- [96] MCAULAY, R. J. and MALPASS, M. L., “Speech enhancement using a soft-decision noise suppression,” *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 28, Apr 1980.
- [97] MCAULAY, R. J. and MALPASS, M. L., “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, Apr 1980.
- [98] MCAULAY, R. J. and QUATIERI, T. F., “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, Aug 1986.
- [99] MELLOR, B. A. and VARGA, A. P., “Noise compensation algorithms for use with hidden Markov model based speech recognition,” in *Proceedings of ICASSP*, (Minneapolis, USA), Sep 1993.
- [100] MOHARIR, G., PATWARDHAN, P., and RAO, P., “Spectral enhancement preprocessing for the HMM coding of noisy speech,” in *ICSLP*, (Denver, Colorado), 2002.
- [101] MOORE, B. C. J., *An Introduction to the psychology of hearing*, ch. 5, pp. 204–208. Academic Press, 1997.
- [102] NAKASHIMA, H., CHISAKI, Y., USAGAWA, T., and EBATA, M., “Spectral subtraction based on statistical criteria of the spectral distribution,” *IEICE Transactions on Fundamentals*, vol. E85-A, Oct 2002.
- [103] NECIOGLU, B., GEORGE, B., SHUTTIC, G., SUNDARAM, R., and PICONE, J., “An interoperability study of speech enhancement and speech recognition systems,” in *ICASSP*, (Salt Lake City, Utah), May 2001.
- [104] NEMER, E., GOUBRAN, R., and MAHMOUD, S., “Speech enhancement using fourth-order cumulants and optimum filters in the subband domain,” *Speech Communication*, vol. 36, 2002.
- [105] NEUMEYER, L. and WEINTRAUB, M., “Probabilistic optimum filtering for robust speech recognition,” in *ICASSP*, (Adelaide, Australia), Apr 1994.
- [106] NIEDERJOHN, R. J. and HEINEN, J. A., “Understanding speech corrupted by noise,” in *IEEE International Conference on Industrial Technology*, (Shanghai, China), Dec 1996.
- [107] OBERLE, S. and KAELIN, A., “HMM-based speech enhancement using pitch period information in voiced speech segments,” in *ISCAS*, (Hong Kong), Jun 1997.
- [108] OPENSHAW, J. P. and MASON, J. S., “On the limitations of cepstral features in noise,” in *ICASSP*, (Adelaide, Austr), Apr 1994.

- [109] PALIWAL, K. K., “On the use of filter-bank energies as features for robust speech recognition,” in *fifth international symposium on signal processing and applications*, (Brisbane, Australia), Aug 1999.
- [110] PEI, S. and TSENG, C., “A comb filter design using fractional-sample delay,” *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, May 1998.
- [111] PELLOM, B. L. and HANSEN, J. H. L., “Automatic segmentation of speech recorded in unknown noisy channel characteristics,” *Speech Communication*, vol. 25, 1998.
- [112] PETEK, B., ANDERSEN, O., and DALSGAARD, P., “On the robust automatic segmentation of spontaneous speech,” in *ICSLP*, (Philadelphia, PA), Oct 1996.
- [113] PICONE, J., <http://www.isip.msstate.edu>.
- [114] PICONE, J. W., “Signal modeling techniques in speech recognition,” *Proceedings of the IEEE*, vol. 81, Sep 1993.
- [115] POHJALAINEN, J., “A new HMM-based approach to broad phonetic classification of speech,” in *EUROSPEECH*, (Geneva, Switzerland), Sep 2003.
- [116] QUACKENBUSH, S. R., BARNWELL, T. P., and CLEMENTS, M. A., *Objective Measures of Speech Quality*. New York, NY: Prentice-Hall, 1988.
- [117] QUATIERI, T. F., BRADY, K., MESSING, D., CAMPBELL, J. P., CAMPBELL, W. M., BRANDSTEIN, M. S., WEINSTEIN, C. J., TARDELLI, J. D., and GATEWOOD, P. D., “Exploiting nonacoustic sensors for speech encoding,” *submitted to IEEE Transactions on Speech and Audio Processing*, 2004.
- [118] QUATIERI, T. F. and DUNN, R. B., “Speech enhancement based on auditory spectral change,” in *ICASSP*, (Orlando, FL), May 2002.
- [119] QUATIERI, T. F., MESSING, D., BRADY, K., CAMPBELL, W. M., CAMPBELL, J. P., BRANDSTEIN, M., WEINSTEIN, C. J., TARDELLI, J. D., and GATEWOOD, P. D., “Exploiting nonacoustic sensors for speech enhancement,” in *Proceedings of the Workshop on Multimodal User Authentication*, (Santa Barbara, CA), Dec 2003.
- [120] RENEVEY, P. and DRYGAJLO, A., “Statistical estimation of unreliable features for robust speech recognition,” in *EUROSPEECH*, (Budapest), Sep 1999.
- [121] SAIJAYRAM, A. K. V. and RAMASUBRAMANIAN, V., “Robust parameters for automatic segmentation of speech,” in *ICASSP*, (Hong Kong), Apr 2003.
- [122] SAMETI, H., SHEIKHZADEH, H., DENG, L., and BRENNAN, R. L., “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, 2001.
- [123] SARIKAYA, R. and HANSEN, J. H. L., “Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition,” in *EUROSPEECH*, (Aalborg, Denmark), Sep 2001.
- [124] SCANLON, M. V., “Acoustic sensor for health status monitoring,” in *Proceedings of IRIS acoustic and seismic sensing*, vol. 2, 1998.

- [125] SEYMOUR, C. W. and NIRANJAN, M., “An HMM-based cepstral-domain speech enhancement system,” in *ICSLP*, (Yokohama, Japan), Sep 1994.
- [126] SKOWRONSKI, M. D. and HARRIS, J. G., “Increased MFCC filter bandwidth for noise-robust phoneme recognition,” in *ICASSP*, (Orlando, FL), May 2002.
- [127] SOON, L. Y., KOH, S. N., and YEO, C. K., “Improved noise suppression filter using self-adaptive estimator of probability of speech absence,” *Signal Processing*, vol. 75, 1999.
- [128] SRINIVASAN, S., SAMUELSSON, J., and KLEIJN, W. B., “Estimation of short-term predictor parameters for coding and enhancement of noisy speech,” in *ICASSP*, (Montreal, Canada), May 2004.
- [129] STOBER, K. and HESS, W., “Additional use of phoneme duration hypotheses in automatic speech segmentation,” in *ICSLP*, (Sydney, Australia), Dec 1998.
- [130] TABRIKIAN, J., DUBNOV, S., and DICKALOV, Y., “Speech enhancement by harmonic modeling via map pitch tracking,” in *ICASSP*, (Orlando, FL), May 2002.
- [131] TCHORZ, J. and KOLLMEIER, B., “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, May 2003.
- [132] THOMSON, D. L. and CHENGALVARAYAN, R., “Extraction methods of voicing feature for robust speech recognition,” *Speech Communication*, vol. 37, 2002.
- [133] TOLBA, H., SELOUANI, S., and O’SHAUGHNESSY, D., “Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for automatic speech recognition using a multi-stream paradigm,” in *ICSLP*, (Denver, Colorado), Sep 2002.
- [134] VAN SANTEN, J. P. H. and SPROAT, R. W., “High-accuracy automatic segmentation,” in *EUROSPEECH*, (Budapest, Hungary), Sep 1999.
- [135] VARGA, A., MOORE, R., BRIDLE, J., PONTING, K., and RUSSELL, M., “Noise compensation algorithms for use with hidden markov model based speech recognition,” in *Proceedings of ICASSP*, (Newyork), Apr 1988.
- [136] VIRAG, N., “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, Mar 1999.
- [137] VISWANATHAN, V. and HENRY, C., “Noise-immune multisensor speech input: formal subjective testing in operational conditions,” in *ICASSP*, vol. 1, (Glasgow, UK), pp. 373 – 376, Oct 1989.
- [138] VISWANATHAN, V., HENRY, C., and DERR, A. G., “Noise-immune speech transduction using multiple sensors,” in *ICASSP*, vol. 2, (Tampa, FL), pp. 711 – 715, Mar 1985.
- [139] VOIERS, W., “The present state of digital vocoding technique: A diagnostic evaluation,” *IEEE Transactions on Audio and Electroacoustics*, vol. 16, 1968.

- [140] VOIERS, W. D., *Speech Intelligibility and Speaker Recognition*, ch. Diagnostic evaluation of speech intelligibility, pp. 374–378. Dowden, Hutchinson, and Ross, Inc., 1977.
- [141] WANG, Y. and YOSHIDA, K., “Speech and noise separations using comb filtering method for high quality speech coding,” in *IEEE Workshop on Speech Coding*, (Ibaraki, JAPAN), Oct 2002.
- [142] WHITEHEAD, P. S., ANDERSON, D. V., and CLEMENTS, M. A., “Adaptive, acoustic noise suppression for speech enhancement,” in *ICME*, (Baltimore, Maryland), Jul 2003.
- [143] WOLFE, P. J. and GODSILL, S. J., “Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement,” in *Proceedings of the 11th IEEE Workshop on Statistical Signal Processing*, 2001.
- [144] WONG, D. and MARKEL, J., “An intelligibility evaluation of several linear prediction vocoder modifications,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, 1978.
- [145] XU, J. and WEI, G., “Noise-robust speech recognition based on difference of power spectrum,” *Electronics Letters*, vol. 36, Jul 2000.
- [146] YAPANEL, U. H. and DHARANIPRAGADA, S., “Perceptual MVDR-based cepstral coefficients (pmccs) for robust speech recognition,” in *ICASSP*, (Hong Kong), Apr 2003.
- [147] Y. ZHENG, LIU, Z., ZHANG, Z., INCLAIR, M., DROPPPO, J., DENG, L., ACERO, A., and HUANG, X., “Multi-sensory microphones for robust speech detection, enhancement and recognition,” in *ICASSP*, (Quebec, Canada), May 2004.
- [148] ZHENG, W. X., “On unbiased parameter estimation of autoregressive signals observed in noise,” in *ICASSP*, (Hong Kong), Apr 2003.
- [149] ZHENG, Y., LIU, Z., ZHANG, Z., SINCLAIR, M., DROPPPO, J., DENG, L., ACERO, A., and HUANG, X., “Air- and bone- conductive integrated microphones for robust speech detection and enhancement,” in *Automatic Speech Recognition and Understanding Workshop*, Dec 2003.
- [150] ZHU, Q. and ALWAN, A., “The effect of additive noise on speech amplitude spectra: A quantitative analysis,” *IEEE Signal Processing Letters*, vol. 9, Sep 2002.
- [151] ZOLNAY, A., SCHULTER, R., and NEY, H., “Extraction methods of voicing feature for robust speech recognition,” in *EUROSPEECH*, (Geneva, Switzerland), Sep 2003.

VITA

Cenk Demiroglu was born in Istanbul, Turkey on March 30, 1977. He attended the Bogazici University, Istanbul, Turkey where he got his Bachelor's degree in Electrical Engineering in 1999. He then attended the University of Nebraska at Lincoln where he completed his Masters of Science in Electrical Engineering in 2001. During this period, he worked as a research assistant on noise-robust image transmission from deep space probes using joint source/channel coding. He continued his phd study at Georgia Institute of Technology in Atlanta, Georgia where he worked as a research assistant. During this time, he also worked as a consultant for CustomspeechUSA Inc., where he was the technical lead of an automatic speech recognition project.

His primary research interest is in noise-robust automatic speech recognition and speech enhancement for improving the quality and intelligibility of noisy speech.