

UNIVERSITY OF CALIFORNIA

Los Angeles

**Inference of Missing or Degraded Data
for Noise Robust Speech Processing**

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Bengt Jonas Borgström

2010

The dissertation of Bengt Jonas Borgström is approved.

Adnan Darwiche

Lieven Vandenberghe

Rick Wesel

Abeer Alwan, Committee Chair

University of California, Los Angeles

2010

till mamma och pappa

TABLE OF CONTENTS

1	Introduction	1
1.1	Noise Robust Speech Processing	1
1.2	Single-Channel Speech Enhancement	3
1.2.1	Statistical Framework	3
1.2.2	Maximum Likelihood Estimation	9
1.2.3	Minimum Mean-Square Error Estimation	10
1.2.4	Maximum a Posteriori Estimation	13
1.2.5	Estimation Under Speech Presence Uncertainty	14
1.2.6	Quantitative Measures for Enhanced Speech Quality	16
1.3	Automatic Speech Recognition	19
1.3.1	Feature Extraction	19
1.3.2	Hidden Markov Models	20
1.3.3	Noise Robustness in ASR	24
1.3.4	Missing Feature Approaches to Noise Robust ASR	24
1.4	Organization of Dissertation	28
I	Short-Time Spectral Amplitude Estimation for Single-Channel Speech Enhancement	29
2	Unified Frameworks for Deriving Short-Time Spectral Amplitude Estimators	30

2.1	A Unified Framework for STSA Estimation Using Generalized Gamma Distributions	30
2.1.1	MAP Estimation	31
2.1.2	MMSE Estimation	32
2.1.3	Estimation Under Speech Presence Uncertainty	34
2.1.4	GGD Shape Parameter Estimation	38
2.1.5	Experimental Results	39
2.2	A Unified Framework for STSA Estimation Assuming Phase Equivalence of Speech and Noise	43
2.2.1	Phase Equivalence and Spectral Subtraction	43
2.2.2	ML Estimation	46
2.2.3	MMSE Estimation	47
2.2.4	MAP Estimation	53
2.2.5	Experimental Results	59
2.3	Speech Enhancement as a Method for Noise Robust ASR	62
3	Exploiting Temporal Correlation in Short-Time Spectral Amplitude Estimation	64
3.1	Statistical Framework	64
3.1.1	Temporal Correlation of Clean STSAs	64
3.1.2	Modeling the Dynamic Behavior of Clean STSAs with Respect to Time	66
3.2	Conditional Probabilities Based on Multiple Observed Spectral Components	68
3.3	Maximum a Posteriori STSA Estimation: the CB-MAP Estimator	71
3.4	Estimation Under Speech Presence Uncertainty: the CB-SPP filter	75

3.5	Experimental Results	78
4	Improved Speech Presence Probabilities Using HMM-Based Inference	81
4.1	Interpreting SPPs as Observations of Channel-Specific 2-State Models	82
4.2	HMM-Based Mask Decoding	83
4.3	Incorporating Future Observations	86
4.4	Complexity Analysis	87
4.5	Experimental Results	88
4.5.1	Accuracy of Improved SPPs	88
4.5.2	Speech Distortion and Noise Leakage	91
II	Front-End Missing Feature Approaches to Noise Robust ASR	93
5	A Statistical Approach to Mel-Domain Mask Estimation	95
5.1	The Mel-Filtered Domain	96
5.2	Mask Estimation	99
5.2.1	Soft-Decision Masks	99
5.2.2	HMM-Based Decoding	100
5.2.3	Binary Masks	102
5.3	Experimental Results	103
6	HMM-Based Reconstruction of Missing Features	106
6.1	Noise Robust Feature Extraction: Log-Spectral Flooring	106
6.2	The Role of HMMs in Spectral Reconstruction	108
6.3	HMM-Based Estimation Methods	114

6.3.1	Utilizing Correlation Across Time	114
6.3.2	Utilizing Correlation Across Frequency Channels	116
6.3.3	Utilizing Correlation Across Time and Across Frequency Channels	119
6.4	Efficient Approximation of HMM-Based Estimation Techniques	119
6.4.1	Downsampling of Statistical Models	119
6.4.2	Adaptation of Statistical Parameters	121
6.4.3	Approximated HMM-Based Estimation	122
6.4.4	Performance and Complexity Analysis	123
6.5	Experimental Results	125
6.5.1	Recognition with Oracle Masks	126
6.5.2	Sensitivity Analysis for κ	127
6.5.3	Recognition with Statistical Masks	128
6.6	Extension to Reconstruction in the Log-Spectral Domain	130
7	Utilizing Compressibility in Reconstructing Spectrographic Data	131
7.1	Signal Recovery from Incomplete Observations	131
7.2	The Compressibility of Spectrographic Speech Data	133
7.3	Reconstruction of Missing Features for Noise Robust Speech Processing .	136
7.3.1	The Proposed Missing Feature Estimation Algorithm	136
7.3.2	Comparisons with Compressive Sensing	138
7.4	Extension to Reconstruction in the Log-Spectral Domain	139
7.5	Experimental Results	140
8	Extension of the Missing Feature Approach to Packet Loss Concealment . .	142

8.1	HMM-Based Estimation	143
8.1.1	Interpreting Speech Parameters as Markov Processes	143
8.1.2	Deriving State-Specific Probabilities	144
8.2	Reducing the Complexity of HMM-Based Estimation	148
8.2.1	Markov Model Downsampling	148
8.2.2	Enforcing Transition Matrix Symmetry	148
8.2.3	Complexity Analysis	150
8.3	Experimental Results	151
9	Summary and Future Work	155
9.1	Short-Time Spectral Amplitude Estimation for Single-Channel Speech Enhancement	155
9.2	Front-End Missing Feature Approaches to Noise Robust ASR	157
9.3	Future Work	159
A	Appendix	160
A.1	Important Statistical Distributions	160
A.1.1	The Rayleigh Distribution	160
A.1.2	The Rice Distribution	160
A.1.3	The χ^2 Distribution	162
A.1.4	The Generalized Gamma Distribution	162
A.2	Special Functions	165
A.2.1	The Gauss Error and Gauss Complementary Error Functions	165
A.2.2	The Gamma and Digamma Functions	166
A.2.3	The Modified Bessel Function of the First Kind	167

A.2.4	The Confluent Hypergeometric Function	168
A.2.5	The Parabolic Cylinder Function	169
A.3	Derivation of Eq. 2.30	169
A.4	Derivation of the Identity in Eq. 2.49	170
A.5	Evaluating the Indeterminant Form of Eq. 2.50	171
References	173

LIST OF FIGURES

1.1	Overview of a remote human-to-human speech communication system utilizing transmitter-end speech enhancement	2
1.2	Overview of an automatic speech recognition system utilizing front-end noise robust processing	2
1.3	An overview of a typical STSA-based speech enhancement system. Note that STFT and ISTFT denote the Short Time Fourier Transform and Inverse Short Time Fourier Transform, respectively.	6
1.4	Gain curve for the Maximum Likelihood STSA estimator from [MM80] .	10
1.5	Gain curves for the Maximum Mean-Square Error STSA estimator from [EM84], for various <i>a priori</i> SNR conditions	12
1.6	Gain curves for the Maximum a Priori STSA estimator from [WG03], for various <i>a priori</i> SNR conditions	14
1.7	Gain curves for the soft-decision speech presence probability filter from [EM84], for various <i>a priori</i> SNR conditions and for $\eta_k=1$	16
1.8	Segmental SNRs for noisy speech from the Noizeus database	18
1.9	COSH distances [JM76] for noisy speech from the Noizeus database . .	19
1.10	Overview of the MFCC computation for ASR front-end feature extraction	20
1.11	General overview of a front-end missing feature ASR system	25
1.12	General overview of a back-end missing feature ASR system	25
2.1	Gain curves for MAP (solid lines) and MMSE (dotted lines) estimators for generalized Gamma distributed speech with $(\zeta=2, \nu=1)$	34
2.2	Gain curves for MAP (solid lines) and MMSE (dotted lines) estimators for generalized Gamma distributed speech with $(\zeta=1, \nu=2)$	35

2.3	Gain curves for the proposed soft-decision speech presence probability filter with $(\zeta=1, \nu=2)$	37
2.4	Estimated shape parameter ν as a function of frequency channel, for the $\zeta=1$ case (top panel) and $\zeta=2$ case (bottom panel)	40
2.5	Subtraction factors, ρ_k , as a function of γ_k and θ_k	44
2.6	Gain curves for the G_{ML} STSA estimator, with $\zeta_n=2$, and for various values of ν_n . Note that for $\nu_n=\infty$, the proposed ML estimator is equivalent to the magnitude spectral subtraction solution from [Bol79].	47
2.7	Gain curves for the \hat{G}_{GMMSE}^1 STSA estimator (solid line) for various values of ξ_k : The Wiener filter [Wie49] (dotted line) is included for comparison.	50
2.8	Gain curves for the \hat{G}_{GEMMSE}^1 STSA estimator for various values of ξ_k .	52
2.9	Gain curves for the $G_{EMMSE}^{(\nu_x)}$ STSA estimator for $\nu_x=1$ (solid line) and $\nu_x=2$ (dotted line)	54
2.10	Gain curves for the $G_{G2MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=1$ (dotted line)	56
2.11	Gain curves for the $G_{G1MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=2.5$ (dotted line)	58
2.12	Gain curves for the G_{RGMAP} (solid line) and G_{REMAP} (dotted line) STSA estimators	59
3.1	An empirical study of power correlation coefficients, obtained from the Noizeus database [HL07]. The top panel illustrates first-order inter-frame power correlation coefficients, $\varrho_k(1)$, as a function of channel index. The bottom panel illustrates inter-frame power correlation coefficients for example frequency channels, as a function of delay, τ	66

3.2	A graphical representation of the proposed model for the dynamic behavior of short-time spectral amplitudes of clean ($X_k(n)$) and observed ($Y_k(n)$) speech.	67
3.3	Gain curves for the proposed correlation-based maximum a posteriori STSA estimator for $\mathcal{T} = \{-1, 0\}$ and for $\gamma_k(n-1)=-5$ dB (solid line), $\gamma_k(n-1)=5$ dB (dotted line), and $\gamma_k(n-1)=10$ dB (dashed line). For illustrative purposes, $\lambda_k=0.95$	74
3.4	Transitional statistics for speech presence probability masks, as defined by Eq. 3.31: As an illustrative example, probabilities are determined with the constraint $a_{11}=a_{00}$	77
3.5	Gain curves for the correlation-based speech presence probability filter for $\mathcal{T}=\{-1, 0\}$, and for $\gamma_k(n-1)=10$ dB (dotted line) and $\gamma_k(n-1)=8$ dB (solid line). For illustrative purposes, $\lambda_k=0.95$ and $\eta_k=0.33$	77
3.6	Improvements in $SSNR_T$ (solid line) and COSH distance (dotted line) for the proposed CB-MAP estimator, as a function of t_1 . Note that for $t_1=0$, the CB-MAP estimator reduces to that proposed in [WG03].	78
4.1	Example probability distribution functions $b_k^0(\tau_k(n))$ and $b_k^1(\tau_k(n))$ for various values of $\xi_k(n)$. For this example, $P(H_1)=0.2$, and $\kappa^i=1.0$	85
4.2	Illustrative examples of SPP masks determined by various methods: Panel (a) provides the clean speech signal "She had your dark suit in greasy wash water all year" spoken by a female. Panel (b) shows the SPP mask determined according to [Coh05] from the corresponding signal corrupted by airport noise at 15 dB SNR. Panels (c) and (d) provide proposed SPP masks according to Eq. 4.8 and Eq. 4.11 ($N_{LA}=2$), respectively.	89

5.1	The Mel-filtering process and the effect on Mel-domain power spectra distributions: The top panel illustrates the Mel-filterbank, as defined in [YKO], for 26 channels. The bottom panel provides the resulting degrees of freedom, k_m , for Mel-channel power spectra χ^2 distributions, as determined empirically by Eq. 5.3.	97
5.2	Examples of Mel-domain mask estimation: Panel a shows the clean utterance "nine one nine six nine five one," from the Aurora-2 database. Panels b and c provide the proposed soft masks, without and with HMM-based decoding, respectively, for the corresponding speech signal degraded by vehicle noise at 5 dB SNR. Panel d illustrates the proposed binary mask obtained by hard thresholding.	103
6.1	Overview of the log-spectral flooring feature extraction process	107
6.2	An illustrative example of log-spectral flooring: The top panel shows the clean and noisy versions of liftered log-spectra corresponding to an active frame of speech (/ε/ from "seven"). The bottom panels shows the same spectra after log-spectral flooring. The noisy speech was degraded by car noise at 10 dB SNR.	108
6.3	An illustrative example of log-spectral flooring: The top panel shows the Mel-filtered spectrogram of the utterance "one two three seven seven four three," from the Aurora-2 database, degraded by car noise at 10 dB SNR. The middle panel shows the spectrogram after liftering. The bottom panel shows the spectrogram after log-spectral flooring.	109

6.4	Example transitional probability matrices, \mathbf{A}_m : For illustrative purposes, log-probabilities are shown. The y-axes correspond to the input state and the x-axes correspond to the output state. Panels a through d refer to Mel-filtered Short-Time Fourier Transform channels with center frequencies 313 Hz, 734 Hz, 1328 Hz, and 2188 Hz, respectively.	112
6.5	An illustrative example of the HMM-based missing feature problem formulation: In this case, the quantizer is comprised of $N=8$ centroids. A series of 3 features are missing. In this figure, the feature at time index n is to be estimated, so that $n_1=2$ and $n_2=2$. Note that in general, n_1 and n_2 are not equal.	114
6.6	The effect of model downsampling on ASR performance: Statistical masks (see Chapter 6.1) were used in series with the \mathbf{FB}_T algorithm for various model resolutions.	124
6.7	Word-accuracy results for \mathbf{FB}_F spectral reconstruction using oracle masks: SS refers to the spectral subtraction-based imputation technique defined in Eq. 6.43, and "none" refers to unprocessed signals. Results were averaged across all noise types in Set A of the Aurora-2 database.	128
7.1	The compressibility of spectrographic speech data: Analysis was performed on the clean word "three" extracted from the Aurora-2 database [Pea00]. The top panel shows the sparse representation of the input spectrographic data in vector form, utilizing the discrete Haar transform (DHT). The bottom panel shows the absolute value of the sparse representation, sorted by magnitude.	134

7.2	Quantitative analysis of the compressibility of spectrographic speech data: The top panel illustrates the MSE distortion resulting from induced sparsity in \mathbf{x}_S , utilizing the discrete Haar transform (DHT). The bottom panel provides word-accuracies corresponding to the recovered Mel-filtered spectra used in the top panel.	135
8.1	Induced complexity of HMM-based estimation of missing features for original method (Eq. 12) and reduced complexity method (Eq. 19)	151
8.2	Reconstructed trajectories for the 1 st LSF in the presence of error bursts. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to HMM-based estimation with $r=7$. Error burst are denoted by horizontal bars.	152
8.3	Weighted LSF distortion [PA93] for estimation of missing LSF features as a function of error rate. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to the proposed HMM-based framework with $r=7$.	153
8.4	The effect of model downsampling on WSD for HMM-based estimation with Eq. 8.11 with a 5% error rate	154
A.1	The Rayleigh distribution for various values of σ^2	161
A.2	The Rice distribution for various combinations of σ^2 and ν	162
A.3	The χ^2 distribution for various degrees of freedom k	163
A.4	The generalized gamma distribution for various shaping parameter pairs: For illustrative purposes, the variance was normalized to unity.	164
A.5	The erf and erfc functions	165
A.6	The Gamma (Γ) and Digamma (Ψ) functions	166
A.7	The modified Bessel function of the first kind for various orders ν	167

A.8	The 0^{th} -order modified Bessel function of the first kind, and its large-value approximation from Eq. A.20	168
A.9	The confluent hypergeometric function for $b=1$ and for various values of ν	169
A.10	The parabolic cylinder function for various orders v	170

LIST OF TABLES

2.1	Segmental SNR scores for MAP and MMSE STSA estimators: Bold entries denote the best score for each metric at each noise condition.	41
2.2	COSH distortion measures for MAP and MMSE STSA estimators. Bold entries denote the best score for each metric at each noise condition.	42
2.3	STSA estimators derived in Sections 2.2.2, 2.2.3, and 2.2.4: Note that particular solutions are obtained by substituting into general solutions those statistical parameters corresponding to desired noise and speech priors.	60
2.4	Segmental SNR scores for selected STSA estimators: Bold entries denote the best score for each metric at each noise condition.	61
2.5	COSH Distortion Measures [JM76] for Selected STSA Estimators. Bold entries denote the best score for each metric at each noise condition. Results were obtained on the Noizeus database [HL07].	62
2.6	Word-accuracy rates for front-end short-time spectral amplitude estimation	62
3.1	Numerical parameters for STSA estimation methods described in Sections 3.3 and 3.4	79
3.2	Segmental SNR scores for proposed STSA estimators	79
3.3	COSH distance [JM76] for proposed STSA estimators	80
4.1	Required operations for proposed SPPs: Numbers of operations are given per frame and per frequency channel. Note that the induced load of the fast Fourier transform (FFT) is not included, but is known to be of order $O(N_{ch} \log(N_{ch}))$	88
4.2	Numerical parameters for proposed SPPs	90

4.3	Mean pointwise Kullback-Leibler (KL) distances for SPP masks from oracle masks, in bits	90
4.4	Speech distortion (SD) and noise leakage (NL) results for proposed SPP masks: Proposed techniques show low SD while providing significantly reduced NL, relative to [Coh05], for most noise conditions.	92
5.1	Parameters utilized during mask estimation	104
5.2	Word-accuracy results for missing feature ASR using various mask estimation techniques, and applying the compressive sensing (CS)-based spectral reconstruction method from Chapter 7	104
6.1	Word-Accuracies for the Aurora-2 database using log-spectral flooring (LS-FLR). Spectral flooring (S-FLR) is included for comparison.	109
6.2	Average operations per frame, as a function of model resolution, r . \mathbf{FB}_T refers to intra-channel HMM-based estimation, \mathbf{FB}_F refers to inter-channel HMM-based estimation, and \mathbf{FB}_{2D} refers to a combination of both. These results were obtained on a single utterance from the Aurora-2 database, with the window update set to 100 Hz, and using 26 Mel channels.	125
6.3	Word-accuracy results for HMM-based spectral reconstruction using oracle masks. "none" refers to unprocessed speech signals. \mathbf{FB}_T refers to estimation along the temporal axis, \mathbf{FB}_F refers to estimation along the frequency axis, and \mathbf{FB}_{2D} refers to a combination of both.	127
6.4	Sensitivity analysis for κ : $\Delta\kappa$ refers to the percent change in κ , \overline{WAcc} refers to the word-accuracy averaged across all noise conditions, and $\Delta\overline{WAcc}$ refers to the change in average word-accuracy relative to that obtained for $\Delta\kappa=0$	129

6.5	Word-accuracy results for HMM-based spectral reconstruction using statistical masks from Chapter 6.1. "none" refers to unprocessed speech signals. \mathbf{FB}_T refers to estimation along the temporal axis, \mathbf{FB}_F refers to estimation along the frequency axis, and \mathbf{FB}_{2D} refers to a combination of both. Results were obtained at a resolution of $r=3$. Bold entries refer to the maximum results for each condition in the average case.	129
6.6	Word-accuracy results for HMM-based spectral reconstruction in the log-spectral domain, using statistical masks from Chapter 6.1	130
7.1	Word-accuracies for the proposed missing feature reconstruction technique, using oracle masks (ORC) and statistical masks (SPP) from Chapter 6.1	140
7.2	Word-Accuracies for the the proposed missing feature reconstruction technique applied in the log-spectral domain, using oracle masks (ORC) and statistical masks (SPP) from Chapter 6.1	141
8.1	Improvements in Peak-SNR (dB), relative to feature repetition, for Estimation of Missing LSF Features as a Function of Error Rate, for $r=7$. Results are Averaged Across 10 Individual LSFs.	153

ACKNOWLEDGMENTS

Firstly, I would like to acknowledge my advisor, Dr. Abeer Alwan, for generously providing me with her time and expertise. It has been a privilege to work with her. As a mentor, she has impacted my life immensely.

I would also like to thank the members of my committee, Dr. Rick Wesel, Dr. Lieven Vandenberghe, and Dr. Adnan Darwiche, for their guidance with my dissertation. Along with Dr. Wesel and Dr. Vandenberghe, the superb instruction of Dr. Villasenor, Dr. Willson, Dr. Laub, and Dr. Sayed has served as an invaluable tool during my research. Furthermore, I would like to thank my SPAPL peers, past and present, for providing an intellectually stimulating research environment.

Finally, I must thank my family for their incredible support throughout my time at UCLA.

This dissertation is supported in part by the NSF.

VITA

1981	Born, Lund, Sweden
2004	BS in Electrical Engineering, UCLA
2005	MS in Electrical Engineering, UCLA
2004-2009	Teaching Assistant Department of Electrical Engineering UCLA
2004-2010	Graduate Student Researcher Department of Electrical Engineering UCLA
2006	Electrical Engineering Department Outstanding Masters Student Award
2006-2009	Consultant Information Systems and Sciences Lab HRL Laboratories
2008	Teaching Fellow Department of Electrical Engineering UCLA
2009-2010	Consultant Office of the CTO Broadcom
2010	PhD in Electrical Engineering, UCLA

PUBLICATIONS

- L. N. Tan, B. Borgstrom, A. Alwan, *Voice Activity Detection Using Harmonic Frequency Components in Likelihood Ratio Test*, ICASSP, pp. 4466-4469, 2010.
- B. J. Borgstrom and A. Alwan, *Improved Speech Presence Probabilities Using HMM-Based Inference, with Applications to Speech Enhancement and ASR*, IEEE Journal of Selected Topics in Signal Processing, to appear, 2010.
- B. J. Borgstrom and A. Alwan, *HMM-Based Reconstruction of Unreliable Spectrographic Data for Noise Robust Speech Recognition*, IEEE Trans. on Audio, Speech, and Language Processing, to appear, 2010.
- P. H. Borgstrom, B. L. Jordan, B. J. Borgstrom, M. J. Stealey, G. S. Sukhatme, M. A. Batalin, and W. J. Kaiser, *NIMS-PL: a Cable-Driven Robot with Self-Calibration Capabilities*, IEEE Trans. on Robotics, Vol. 25, No. 5, pp. 1005-1015, 2009.
- V. Mitra, B. J. Borgstrom, C. Espy-Wilson, and A. Alwan, *A Noise-type and level-dependent MPO-based speech enhancement architecture*, pp. 2751-2754, Interspeech 2009.
- B. J. Borgstrom and A. Alwan, *Missing Feature Imputation of Log-Spectral Data For Noise Robust ASR*, Workshop on DSP in Mobile and Vehicular Systems, 2009.
- B. J. Borgstrom and A. Alwan, *Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR*, IEEE Signal Processing

Letters, Vol. 16, Issue 5, pp. 398-401, 2009.

B. J. Borgstrom and A. Alwan, *HMM-Based Estimation of Unreliable Spectral Components for Noise Robust Speech Recognition*, Interspeech 2008, pp. 1769-1772.

B. J. Borgstrom and A. Alwan, *An Efficient Approximation of the Forward-Backward Algorithm to Deal With Packet Loss, With Applications to Remote Speech Recognition*, ICASSP 2008, pp. 4425-4428.

B. J. Borgstrom and A. Alwan, *A Low Complexity Parabolic Lip Contour Model with Speaker Normalization for High-Level Feature Extraction in Noise-Robust Audio-Visual Speech Recognition*, IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans, Vol. 38, No. 6, pp. 1273-1280, 2008.

B. J. Borgstrom, A. Bernard, and A. Alwan, *Error Recovery - Channel Coding and Packetization*, Chapter 8 in *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, Springer-Verlag. Editors: Z.-H. Tan and B. Lindberg, pp. 163-185, 2008

B. J. Borgstrom and A. Alwan, *A Packetization and Variable Bitrate Interframe Compression Scheme For Vector Quantizer-Based Distributed Speech Recognition*, Proceedings of Interspeech 2007, pp. 578-581, Belgium.

B. J. Borgstrom, M. van der Schaar, and A. Alwan, *Rate Allocation for Noncollaborative Multiuser Speech Communication Systems Based on Bargaining Theory*, IEEE Trans. on Audio, Speech, and Language, Processing, Vol. 15, No. 4, pp. 1156-1166, 2007.

- P. Oh, P. Borgstrom, H. Witkiewicz, Y. Lil, B. J. Borgstrom, A. Chrastina, K. Iwata, K. R. Zinn, R. Baldwin, J. E. Testa, and J. E. Schnitzer, *Live Dynamic Imaging of Calveolae Pumping Targeted Antibody Rapidly and Specifically Across Endothelium in the Lung*, Nature Biotech, 25, pp. 237-337, 2007.
- B. J. Borgstrom, M. van der Schaar, A. Alwan, *Bargaining-Based Rate Allocation for Non-Collaborative Multi-User Speech Communication Systems*, SiMPE workshop, 2006.
- J. Xue, B. J. Borgstrom, J. Jiang, L. Bernstein, A. Alwan, *Acoustically-driven Talking Face Synthesis Using Dynamic Bayesian Networks*, Proceedings of IEEE ICME 2006, pp. 1165-1168.

ABSTRACT OF THE DISSERTATION

Inference of Missing or Degraded Data
for Noise Robust Speech Processing

by

Bengt Jonas Borgström

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2010

Professor Abeer Alwan, Chair

In real world speech processing systems, speech signals are often corrupted by background acoustic noise or reverberation. Additionally, for systems which involve transmission of speech data over error-prone communication channels, signals may suffer from packet loss. This dissertation addresses two general frameworks for which compensation of corruptive acoustic noise and channel errors can benefit performance, namely remote speech communication and automatic speech recognition.

In the case of ASR, front-end missing feature (MF) spectral reconstruction is explored. Two solutions are offered, the first of which uses HMM-based processing and accounts for temporal and/or frequency correlation. The second exploits the sparsity of spectrographic speech data to formulate the reconstruction problem as a linear program. Each approach is successfully applied in both the Mel-filtered and log Mel-filtered domains. Finally, a statistical approach to Mel-domain mask estimation is proposed, which is used to differentiate between reliable and unreliable time-frequency components. Theory de-

veloped for missing feature reconstruction is extended to the application of packet loss concealment during the transmission of speech features over an error-prone channel.

In the case of single-channel speech enhancement, statistical model-based methods are studied. A unified framework is presented for deriving short-time spectral amplitude (STSA) estimators which assume generalized Gamma-distributed speech priors. Additionally, a unified framework is proposed for developing STSA estimators which assume phase equivalence of speech and noise components. Finally, the role of temporal correlation in statistical speech enhancement is explored, resulting in a novel correlation-based STSA estimator.

CHAPTER 1

Introduction

1.1 Noise Robust Speech Processing

In real world speech processing systems, speech signals are often corrupted by background acoustic noise or reverberation. Additionally, for systems which involve transmission of speech data over error-prone communication channels, signals may suffer from packet loss. This dissertation addresses two general frameworks for which compensation of corruptive acoustic noise and channel errors can benefit performance, namely remote speech communication and automatic speech recognition.

During transmission of speech for the purpose of human-to-human communication, acoustic degradation experienced at the transmitter will result in decreased perceptual quality at the receiver. To combat this, the observed signal can be enhanced prior to transmission. Furthermore, packet loss experienced during the transmission of speech information will result in decreased signal quality corresponding to missing segments. The perceived effect of dropped packets at the receiver can be minimized by applying packet loss concealment (PLC) to estimate missing speech features. Figure 1.1 illustrates a general overview of a remote human-to-human speech communication system utilizing transmitter-end speech enhancement and receiver-end PLC.

For human-to-machine interaction, noise may distort discriminative speech information vital for automatic recognition. Noise robustness for ASR can be approached in front-end signal processing, or back-end recognition ([Ace93], [Gon95]). The former generally

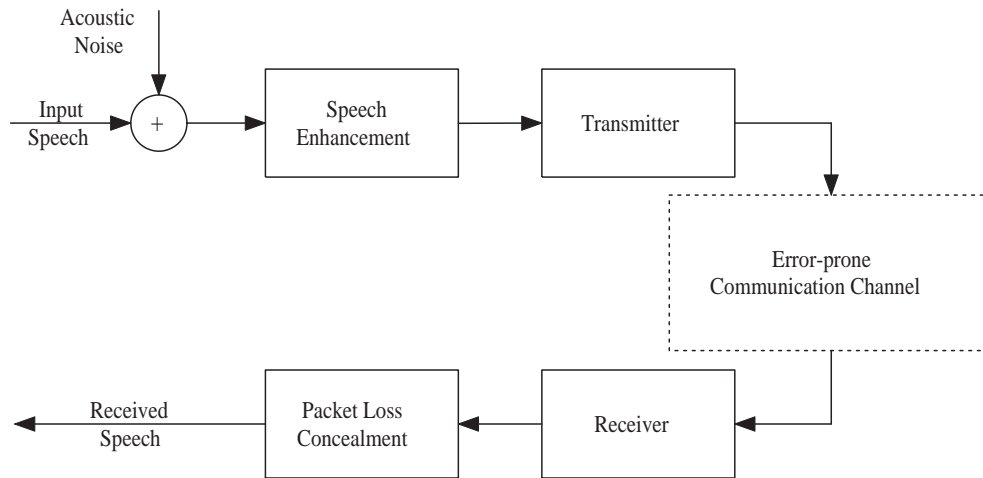


Figure 1.1: Overview of a remote human-to-human speech communication system utilizing transmitter-end speech enhancement

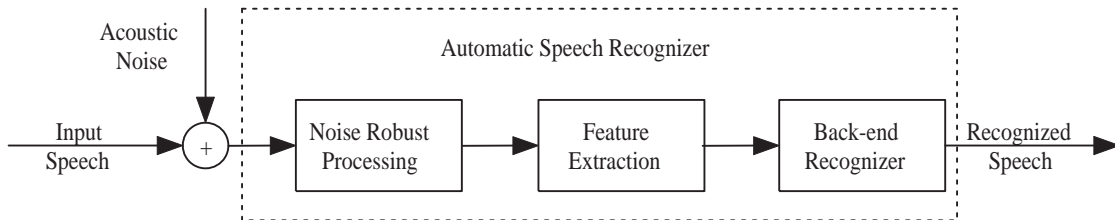


Figure 1.2: Overview of an automatic speech recognition system utilizing front-end noise robust processing

proves more computationally efficient, and avoids adaptation of pre-trained acoustic models. Figure 1.2 illustrates a general overview of an automatic speech recognition (ASR) system utilizing front-end noise robust processing.

The two tasks of increasing perceptual quality of speech signals, and improving ASR performance exhibit many fundamental differences. However, at the heart of each lies the general problem of inferring missing or degraded speech data, where signal ambiguity is due to acoustic or channel noise.

1.2 Single-Channel Speech Enhancement

Single-channel speech enhancement is the task of estimating the clean underlying speech signal given a degraded signal observed at a single sensor. In this section, an overview of single-channel speech enhancement is provided, along with the notation and statistical framework which will be used throughout this dissertation.¹

1.2.1 Statistical Framework

Assuming an additive noise model, an observed speech signal can be expressed in the time domain as

$$y(t) = x(t) + n(t) \quad (1.1)$$

where $y(t)$, $x(t)$, and $n(t)$ represent the observed speech, clean speech, and noise, respectively, and t represents the discrete sample number. As is typical in speech analysis, the short-time Fourier transform (STFT) [Qua01] is used to reveal local frequency content of the input signal, as it varies with time. This requires a time-limited analysis window $w(t)$, of duration N_w , to be applied to the input signal with a period of N_h samples

$$Y_k(n) = \sum_{t=-\infty}^{\infty} y(t) w(t - nN_h) \exp\left(-\frac{j2\pi kt}{N_w}\right) \quad (1.2)$$

leading to

$$Y_k(n) = X_k(n) + N_k(n), \quad (1.3)$$

where Y_k , X_k , and N_k represent the complex spectral coefficients of observed speech,

¹Although single- and multi-channel speech enhancement systems both exist, we assume a framework that is only applicable to the former. Thus, mention of the term "single-channel" will generally be excluded.

clean speech, and noise, respectively, and where k denotes channel index. Note that Eq. 1.3 includes the time index n . For the sake of notation, time indices will be excluded unless required to differentiate between inter-frame components. Spectral coefficients can be decomposed into magnitude and phase components

$$Y_k = R_k \exp(j\varphi_k) \quad (1.4)$$

$$X_k = A_k \exp(j\alpha_k) \quad (1.5)$$

$$N_k = D_k \exp(j\psi_k) \quad (1.6)$$

Here, R_k , A_k , and D_k denote the short-time spectral amplitudes (STSAs) of observed speech, clean speech, and noise, respectively, and φ_k , α_k , and ψ_k are the corresponding phases.

Statistical analysis of speech and noise signals relies heavily on channel-dependent signal-to-noise (SNR) ratios. As in [MM80], the *a priori* and *a posteriori* SNRs, ξ_k and γ_k , respectively, are defined as

$$\begin{aligned} \xi_k &= \frac{E[A_k^2]}{E[D_k^2]} = \frac{\sigma_{X,k}^2}{\sigma_{N,k}^2} \\ \gamma_k &= \frac{R_k^2}{E[D_k^2]} = \frac{R_k^2}{\sigma_{N,k}^2}. \end{aligned} \quad (1.7)$$

Here, $\sigma_{X,k}^2$ and $\sigma_{N,k}^2$ represent the noncentral second moments of speech and noise spectral amplitudes, respectively. The *a posteriori* SNR is a function of the instantaneous observation R_k , and can therefore be interpreted as a local measure. In fact, the value γ_k is often referred to as the *instantaneous SNR*. The *a priori* SNR, which can be interpreted as a global measure, relies on second-order statistics of the hidden clean speech signal, and

must be approximated. A spectral subtraction approach yields [Bol79]

$$\xi_k = \frac{\max\{R_k^2 - \sigma_{N,k}^2, 0\}}{\sigma_{N,k}^2} = \max\{\gamma_k - 1, 0\} \quad (1.8)$$

In [EM84], Ephraim and Malah propose the *decision-directed* (DD) approach which includes smoothing along with a non-linear "decision" term

$$\xi_k(n) = \kappa \frac{\hat{A}_k^2(n-1)}{\sigma_{N,k}^2} + (1 - \kappa) \max[\gamma_k(n-1) - 1, 0] \quad (1.9)$$

where \hat{A}_k is the estimated speech amplitude and $0 \ll \kappa < 1$. More advanced methods by which to estimate the second-order statistics of speech include the ARCH and GARCH models [Coh04].

A common framework for STSA speech enhancement involves determining the frequency representation of a windowed speech segment, yielding the complex-valued Y_k . Given the observed spectrum R_k , the clean speech spectral amplitude, \hat{A}_k is inferred. The noisy phase φ_k is then applied to the estimated spectrum, from which the enhanced time signal can be synthesized. Figure 1.3 provides an overview of a typical STSA-based speech enhancement system. Note that use of the noisy phase during enhancement is motivated by the insensitivity of human hearing to phase information [WL82]. Additionally, the noisy phase can be shown to be the maximum likelihood estimate of the clean phase α_k [EM84].

This work focuses on estimation of the clean spectrum \hat{A}_k based on the observed signal, R_k . The enhancement operation can generally be expressed conveniently as a gain function

$$G(\xi_k, \gamma_k) = \frac{\hat{A}_k}{R_k} \quad (1.10)$$

and is typically a function of *a priori* and *a posteriori* SNRs. Although various speech

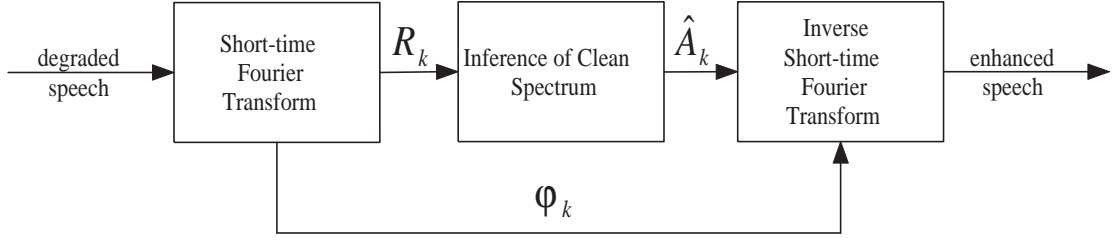


Figure 1.3: An overview of a typical STSA-based speech enhancement system. Note that STFT and ISTFT denote the Short Time Fourier Transform and Inverse Short Time Fourier Transform, respectively.

enhancement approaches exist, such as spectral subtraction and subspace techniques, this work explores statistical model-based methods by which to infer $G(\xi_k, \gamma_k)$. Note that Eq. 1.10 assumes statistical independence of successive STSAs. Inter-frame correlation has not been explored explicitly in the context of statistical speech enhancement.

Statistical analysis for noise robust speech processing requires stochastic models for speech and noise processes. Due in part to the resulting mathematical efficiency, complex spectral coefficients of noise are generally modeled by a Gaussian random process with a diagonal covariance matrix ([MM80], [EM84], [EM85])

$$E \begin{bmatrix} \Re\{N_k\} \\ \Im\{N_k\} \end{bmatrix} \begin{bmatrix} \Re\{N_k\} \\ \Im\{N_k\} \end{bmatrix}^* = \begin{bmatrix} \sigma_{N,k}^2/2 & 0 \\ 0 & \sigma_{N,k}^2/2 \end{bmatrix} \quad (1.11)$$

where $*$ denotes the conjugate transpose, and \Re and \Im refer to real and imaginary components, respectively. This leads to the spectral coefficient distribution

$$p(N_k) = \frac{1}{\pi \sigma_{N,k}^2} \exp\left(-\frac{D_k^2}{\sigma_{N,k}^2}\right) \quad (1.12)$$

Gaussian *a priori* processes in the complex spectral coefficient domain correspond to phase-invariant distributions in the spectral amplitude domain, which can be obtained via

marginalization

$$p(D_k) = \int_0^{2\pi} \frac{D_k}{\pi\sigma_{N,k}^2} \exp\left(-\frac{D_k^2}{\sigma_{N,k}^2}\right) \partial\varphi = \frac{2D_k}{\sigma_{N,k}^2} \exp\left(-\frac{D_k^2}{\sigma_{N,k}^2}\right) \quad (1.13)$$

Note that Eq. 1.13 follows a Rayleigh distribution (see Appendix A.1.1). Furthermore, assuming a Gaussian noise model allows the conditional distribution of observed spectral coefficients given the clean speech spectral amplitude to be elegantly reduced to a closed form expression involving the modified Bessel function [MM80]

$$\begin{aligned} p(Y_k|A_k) &= \int_0^{2\pi} \frac{1}{\pi\sigma_{N,k}^2} \exp\left(-\frac{|Y_k - A_k e^{j\alpha_k}|^2}{\sigma_{N,k}^2}\right) \partial\alpha_k \\ &= \frac{1}{2\pi^2\sigma_{N,k}^2} \exp\left(-\frac{R_k^2 + A_k^2}{\sigma_{N,k}^2}\right) \int_0^{2\pi} \exp\left(\frac{2A_k\Re\{Y_k e^{-j\alpha_k}\}}{\sigma_{N,k}^2}\right) \partial\alpha_k \\ &= \frac{1}{\pi\sigma_{N,k}^2} I_0\left(\frac{2A_k R_k}{\sigma_{N,k}^2}\right) \exp\left(-\frac{R_k^2 + A_k^2}{\sigma_{N,k}^2}\right), \end{aligned} \quad (1.14)$$

where $I_v(\cdot)$ denotes the v^{th} -order modified Bessel function of the first kind (see Appendix A.2.3). Finally, the conditional probability of the observed spectral amplitude can be derived by marginalization with respect to φ_k

$$\begin{aligned} p(R_k|A_k) &= \int_0^{2\pi} \frac{R_k}{\pi\sigma_{N,k}^2} I_0\left(\frac{2A_k R_k}{\sigma_{N,k}^2}\right) \exp\left(-\frac{R_k^2 + A_k^2}{\sigma_{N,k}^2}\right) \partial\varphi_k \\ &= \frac{2R_k}{\sigma_{N,k}^2} I_0\left(\frac{2A_k R_k}{\sigma_{N,k}^2}\right) \exp\left(-\frac{R_k^2 + A_k^2}{\sigma_{N,k}^2}\right) \end{aligned} \quad (1.15)$$

which follows a Rice distribution (see Appendix A.1.2).

Traditionally, Gaussian models were assumed for the clean speech process ([EM84],

[EM85]). Recently, super-Gaussian *a priori* models have been considered, since such models may more accurately describe the narrow peak and prominent tail exhibited by empirical distributions ([Mar05], [CL07]). However, super-Gaussian processes in the spectral coefficient domain correspond to mathematically complicated, generally phase-dependent distributions in the spectral amplitude domain. To circumvent these issues, Lotter and Vary propose a phase-invariant super-Gaussian STSA model fit to randomly generated data [LV05]. In [EHH07], the authors model STSAs with the generalized Gamma distribution (GGD) (see Appendix A.1.4), which allows flexibility in capturing the underlying statistical behavior of speech. To the best of the authors' knowledge, no study exists which provides a uniform framework to STSA estimation using GGD priors, which extends to topics such as speech presence uncertainty and shape parameter estimation.

Single-channel speech enhancement techniques can be grouped into three main approaches: spectral subtraction methods, subspace methods, and statistical model-based approaches [Loi07]. Spectral subtraction builds upon the assumption of phase equivalence of speech and noise. To improve upon the traditional magnitude spectral subtractor (MSS) of [Bol79], variations have been proposed to minimize the distortion of perceptually important speech components ([BSM79], [LB92]). Such methods, however, suffer from their heuristic nature. To the best of the authors' knowledge, the application of statistical optimization criteria within the framework of spectral subtraction has not been explored. In Sections 1.2.2-1.2.4, a review of maximum likelihood (ML), minimum mean-square error (MMSE), and maximum a posteriori (MAP) estimation is provided, and traditional STSA solutions are discussed. In Section 1.2.5, the notion of speech presence uncertainty is reviewed.

1.2.2 Maximum Likelihood Estimation

Maximum likelihood (ML) estimation offers an efficient statistical method when the *a priori* distribution of the target signal is unknown. When applied to the current task of STSA estimation, the maximum likelihood solution is expressed as

$$\hat{A}_k = \arg \max_{A_k} p(Y_k|A_k) \quad (1.16)$$

If a global solution exists, the ML estimator can be determined by

$$\begin{aligned} \hat{A}_k = A_k \text{ such that } (i) \quad & \frac{\partial}{\partial A_k} \mathfrak{F}\{p(Y_k|A_k)\} = 0 \\ (ii) \quad & \frac{\partial^2}{\partial A_k^2} \mathfrak{F}\{p(Y_k|A_k)\} < 0 \end{aligned} \quad (1.17)$$

where \mathfrak{F} is monotonically increasing. Since exponential functions are generally present in statistical distributions of speech and noise, efficient derivation is often achieved with $\mathfrak{F}=\log$.

In [MM80], McAulay and Malpass derive the ML STSA estimator assuming a deterministic clean speech signal. Substituting Eq. 1.14 and Eq. A.20 into Eq. 1.17, constraint (i) leads to

$$\frac{2A_k^2}{\sigma_{N,k}^2} - \frac{2R_k A_k}{\sigma_{N,k}^2} + \frac{1}{2} = 0 \quad (1.18)$$

Applying the quadratic equation to Eq. 1.18 yields

$$G(\xi_k, \gamma_k) = \frac{1}{2} \left(1 + \sqrt{1 - \frac{1}{\gamma_k}} \right) \quad (1.19)$$

It can be shown that Eq. 1.17, constraint (ii) holds for $A_k > 0$, thereby verifying the

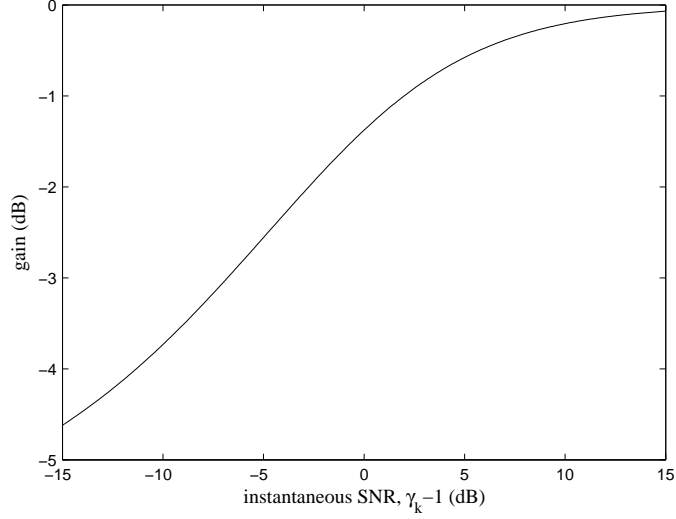


Figure 1.4: Gain curve for the Maximum Likelihood STSA estimator from [MM80]

solution of Eq. 1.19 to be valid. Figure 1.4 illustrates the gain curve for the Maximum Likelihood STSA estimator from [MM80]. As follows intuitively, attenuation is shown to increase as the instantaneous SNR decreases. Note that Eq. 1.19 is independent of ξ_k , which is a result of not accounting for the *a priori* distribution of X_k .

1.2.3 Minimum Mean-Square Error Estimation

When *a priori* distributions are accessible, the minimum mean-square error estimate of the clean speech spectral component can be obtained by minimizing the minimum square-error (MSE) Bayes' risk

$$C = \int_0^\infty (A_k - \hat{A}_k)^2 p(A_k|Y_k) \partial A_k \quad (1.20)$$

Differentiating Eq. 1.20 with respect to A_k , and equating to zero, leads the MMSE STSA estimator to be expressed as

$$G(\xi_k, \gamma_k) = \frac{1}{R_k} \int_0^\infty A_k p(A|Y_k) \partial A_k \quad (1.21)$$

Applying Bayes' rule results in

$$G(\xi_k, \gamma_k) = \frac{1}{R_k} \frac{\int_0^\infty A_k p(Y_k|A_k) p(A_k) \partial A_k}{\int_0^\infty p(Y_k|A_k) p(A_k) \partial A_k} \quad (1.22)$$

If speech spectral coefficients are assumed Gaussian, the spectral amplitude A_k is Rayleigh-distributed

$$p(A_k) = \frac{2A_k}{\sigma_{X,k}^2} \exp\left(-\frac{A_k^2}{\sigma_{X,k}^2}\right) \quad (1.23)$$

Applying Eqs. 1.14 and 1.23 to the MMSE solution from Eq. 1.22 yields

$$G(\xi_k, \gamma_k) = \frac{1}{R_k} \frac{\int_0^\infty \frac{2(1+\xi_k)A_k^2}{\xi_k \sigma_{N,k}^2} I_0\left(\frac{2A_k R_k}{\sigma_{N,k}^2}\right) \exp\left(-\frac{(1+\xi_k)^2 A_k^2 + \xi_k^2 R_k^2}{\xi_k (1+\xi_k) \sigma_{N,k}^2}\right) \partial A_k}{\int_0^\infty \frac{2(1+\xi_k)A_k}{\xi_k \sigma_{N,k}^2} I_0\left(\frac{2A_k R_k}{\sigma_{N,k}^2}\right) \exp\left(-\frac{(1+\xi_k)^2 A_k^2 + \xi_k^2 R_k^2}{\xi_k (1+\xi_k) \sigma_{N,k}^2}\right) \partial A_k} \quad (1.24)$$

It can be observed that the numerator of Eq. 1.24 is equivalent to the first noncentral moment of a Rice distribution with parameters $\xi_k R_k / (1 + \xi_k)$ and $0.5 \xi_k \sigma_{N,k}^2 / (1 + \xi_k)$ (see Appendix A.1.2). Additionally, the integral in the denominator can be observed to be the integration of a Rice distribution over its complete range, thereby reducing to unity, leading to the MMSE solution from [EM84]

$$G(\xi_k, \gamma_k) = \Gamma(1.5) \sqrt{\frac{\xi_k}{\gamma_k (1 + \xi_k)}} {}_1F_1\left(-\frac{1}{2}; 1; -\frac{\xi_k \gamma_k}{1 + \xi_k}\right) \quad (1.25)$$

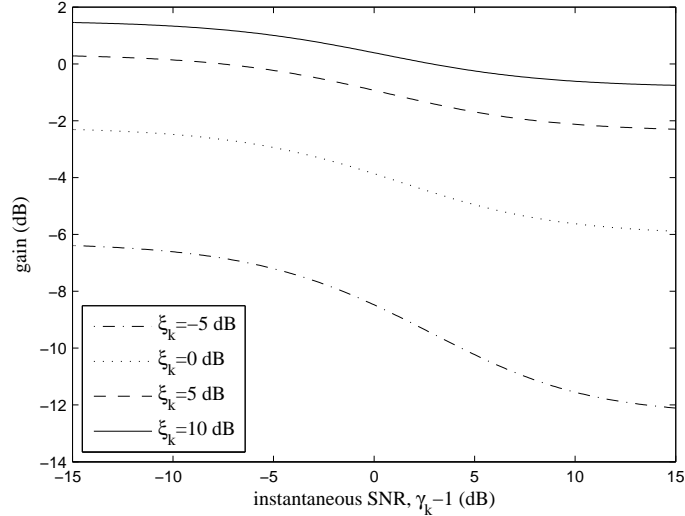


Figure 1.5: Gain curves for the Maximum Mean-Square Error STSA estimator from [EM84], for various *a priori* SNR conditions

where ${}_1F_1$ is the confluent hypergeometric function (see Appendix A.2.4). Figure 1.5 illustrates gain curves for the MMSE STSA estimator from Eq. 1.25, for various values of ξ_k . Note that attenuation increases for less favorable *a priori* SNR conditions, which follows intuitively. However, as opposed to the ML solution from Eq. 1.19, attenuation decays as the instantaneous SNR decreases, especially for low ξ_k . As discussed in [EM84], this characteristic is a result of the statistical estimator compromising between *a priori* knowledge provided by ξ_k , and new information introduced by γ_k . The MMSE solution of Eq. 1.25 is presented by Ephraim and Malah in [EM84]. Several studies have since proposed perceptually motivated STSA estimators by manipulating the MMSE cost function (Eq. 1.20) to include some aspect relevant to the human auditory system. In [EM85], Ephraim and Malah derive a solution which minimizes the log-domain MSE error. In [Loi05], Loizou proposes a family of perceptually-motivated MMSE estimators. In [PC08], Plourde and Champaign provide a generalized framework for auditory-based MMSE STSA estimators.

1.2.4 Maximum a Posteriori Estimation

MMSE estimation may lead to mathematically complex or irreducible solutions, which may not be desirable for speech applications. In [WG03], Wolfe and Godsil present an efficient alternative to Eq. 1.25 using the maximum a posteriori criterion. When applied to the problem of short-time spectral amplitude estimation, the MAP solution is expressed as

$$\hat{A}_k = \arg \max_{A_k} \mathfrak{F} \{p(Y_k|A_k) p(A_k)\}, \quad (1.26)$$

If a solution exists, the solution can be found through differentiation

$$\begin{aligned} \hat{A}_k = A_k \text{ such that } (i) \quad & \frac{\partial}{\partial A_k} \mathfrak{F} \{p(Y_k|A_k) p(A_k)\} = 0 \\ (ii) \quad & \frac{\partial^2}{\partial^2 A_k} \mathfrak{F} \{p(Y_k|A_k) p(A_k)\} < 0 \end{aligned} \quad (1.27)$$

Assuming clean speech follows a Gaussian *a priori* distribution, using $\mathfrak{F}=\log$, and applying Eq. A.20, leads Eq. 1.27, constraint (i) to reduce to

$$A_k^2 - \frac{\xi_k R_k A_k}{1 + \xi_k} + \frac{\xi_k \sigma_{N,k}^2}{4(1 + \xi_k)} = 0 \quad (1.28)$$

Applying the quadratic equation yields

$$G(\xi_k, \gamma_k) = \frac{\xi_k + \sqrt{\xi_k^2 + (1 + \xi_k) \xi_k / \gamma_k}}{2(1 + \xi_k)} \quad (1.29)$$

It is easy to show that Eq. 2.4, constraint (ii) holds for $A_k > 0$, showing the MAP solution of Eq. 1.29 to be valid. Figure 1.6 illustrates gain curves for the maximum a posteriori STSA estimator from [WG03], for various *a priori* SNR conditions. Note that

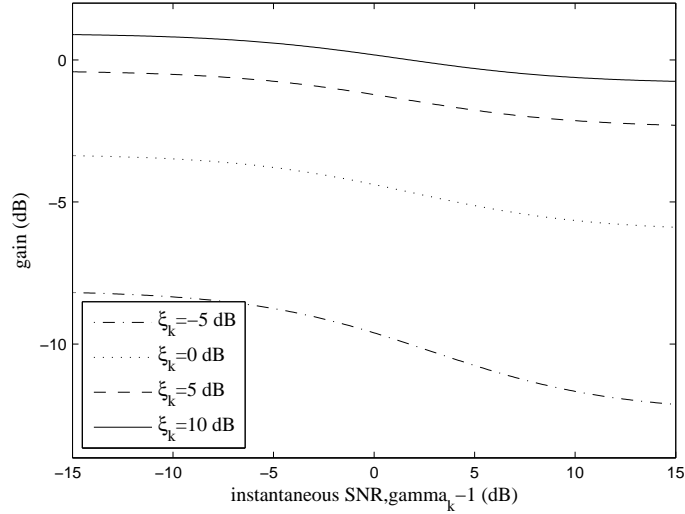


Figure 1.6: Gain curves for the Maximum a Priori STSA estimator from [WG03], for various *a priori* SNR conditions

the behavior of the MAP estimator from [WG03] is similar to that of the MMSE estimator [EM84] illustrated in Figure 1.5.

Besides resulting in computationally efficient solutions to STSA estimation, the MAP criterion provides the mathematical flexibility to apply a variety of statistical models for *a priori* speech and noise processes. In [LV03], Lotter and Vary present MAP estimators which assume a Gamma distribution for clean speech STSAs. In [DTI05] and [PSS04], generalized versions of the Gamma distribution are used.

1.2.5 Estimation Under Speech Presence Uncertainty

Short-time spectral amplitude estimators operate under the assumption of speech presence throughout time-frequency components. This, however, is untrue due in part to temporal speech pauses, and artifacts of speech production such as vocal tract zeros and harmonic valleys. Building upon decision theory developed for the detection of general signals in noise [Esp68], McAulay and Malpass propose STSA estimation under speech presence uncertainty by implementing a soft-decision speech presence probability (SPP)

filter. In [MM80], a two-state model for speech activity is assumed, wherein H_1 denotes the superposition of speech and noise, and H_0 denotes signal comprised solely of noise [MM80]

$$q_k = \begin{cases} H_0 \Rightarrow Y_k = N_k \\ H_1 \Rightarrow Y_k = A_k + N_k \end{cases} \quad (1.30)$$

A minimum mean-square error estimate reveals a combined estimator which considers speech presence uncertainty

$$\tilde{G}(\xi_k, \gamma_k) = \frac{\Lambda(\xi_k, \gamma_k)}{1 + \Lambda(\xi_k, \gamma_k)} G(\xi_k, \gamma_k), \quad (1.31)$$

where $G(\xi_k, \gamma_k)$ is the original estimator, and $\Lambda(\xi_k, \gamma_k)$ refers to the generalized likelihood ratio (GLR)

$$\Lambda(\xi_k, \gamma_k) = \frac{p(q_k = H_1 | Y_k)}{p(q_k = H_0 | Y_k)} = \eta_k \frac{p(Y_k | q_k = H_1)}{p(Y_k | q_k = H_0)}. \quad (1.32)$$

Here, η_k is the ratio of steady-state probabilities

$$\eta_k = \frac{P(q_k = H_1)}{P(q_k = H_0)} \quad (1.33)$$

In [MM80] the clean speech process is modeled as a deterministic signal during the derivation of the GLR. In [EM84], Ephraim and Malah instead assume a Gaussian *a priori* model, leading to

$$\Lambda(\xi_k, \gamma_k) = \frac{\eta_k}{1 + \xi_k} \exp\left(\frac{\xi_k \gamma_k}{1 + \xi_k}\right) \quad (1.34)$$

Figure 1.7 illustrates gain curves for the soft-decision speech presence probability filter from [EM84], for various *a priori* SNR conditions. Note that for favorable conditions

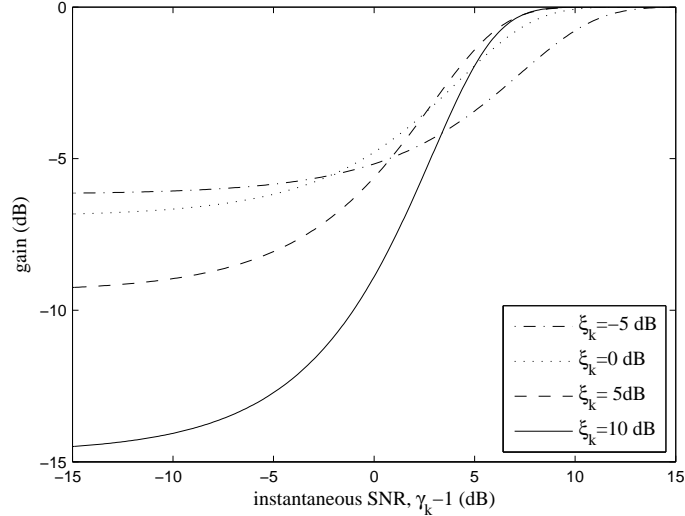


Figure 1.7: Gain curves for the soft-decision speech presence probability filter from [EM84], for various *a priori* SNR conditions and for $\eta_k=1$

(eg. $\xi_k=10$ dB), negligible attenuation is applied for large instantaneous SNRs, whereas high attenuation is applied for small instantaneous SNRs. In such cases, the H_1 hypothesis dominates Eq. 1.32 as γ_k increases, and the H_0 hypothesis dominates as γ_k decreases. Conversely, for poor conditions (eg. $\xi_k=-5$ dB), the uncertainty introduced by corruptive acoustic noise leads to neither hypothesis being heavily favored. Note that the GLR expression from Eq. 1.32 is dependent on the *a priori* distribution of speech. In [CL07] and [Mar05], GLRs are presented for super-Gaussian models.

1.2.6 Quantitative Measures for Enhanced Speech Quality

In order to assess the success of speech enhancement, algorithms are applied to databases comprised of corrupted speech utterances, which are generally created by synthetically adding noise to clean speech. Although such mixing may be considered unnatural, it simplifies the recording process and allows for noisy speech signals at specific signal-to-noise (SNR) ratios. Additionally, it provides *oracle* information of the clean signal during experimental analysis. Many speech enhancement studies utilize the TIMIT database

[GLJ93] during noise mixing, which is comprised of phonetically balanced clean speech utterances. Various noise signal databases include the Noisex [VS93] and Aurora [Pea00] sets.

The Noizeus database [HL07] was introduced to provide a standard set of degraded speech utterances. It is comprised of 30 TIMIT utterances with 8 types of non-stationary additive noise from the Aurora set, namely restaurant, car, train, babble, station, exhibition, street, and airport. To simulate the frequency response of telephone headsets, degraded signals are filtered by the modified Intermediate Reference Systems (MIR) filter, defined in [P800].

Various quantitative metrics exist to assess the quality of enhanced speech. A common measure is segmental SNR between the enhanced signal and the oracle clean signal. In order to analyze the effect of enhancement on active and inactive speech segments separately, [LV05] proposes to use the SSNR of active speech frames ($SSNR_S$) and inactive speech frames ($SSNR_N$). For an enhanced time-domain speech signal $\hat{x}(t)$ and clean reference signal $x(t)$, the global SSNR is defined as

$$SSNR_T(x, \hat{x}) = \frac{1}{N_f} \sum_{k=1}^{N_f} \left[10 \log_{10} \left(\frac{\sum_{i=1}^{N_w} x^2(i + kN_h)}{\sum_{i=1}^{N_w} (x(i + kN_h) - \hat{x}(i + kN_h))^2} \right) \right] \quad (1.35)$$

where N_w is the length of the analysis window, N_h is the shift in samples for each successive window, and N_f is the number of frames analyzed. The $SSNR_S$ is determined similarly to Eq. 1.35, although only computed for speech frames wherein the clean signal energy is ≥ -30 dB. Conversely, the $SSNR_N$ is computed only for frames exhibiting energy < -30 dB. SNR-related measures are often reported in terms of improvement over the unprocessed signal

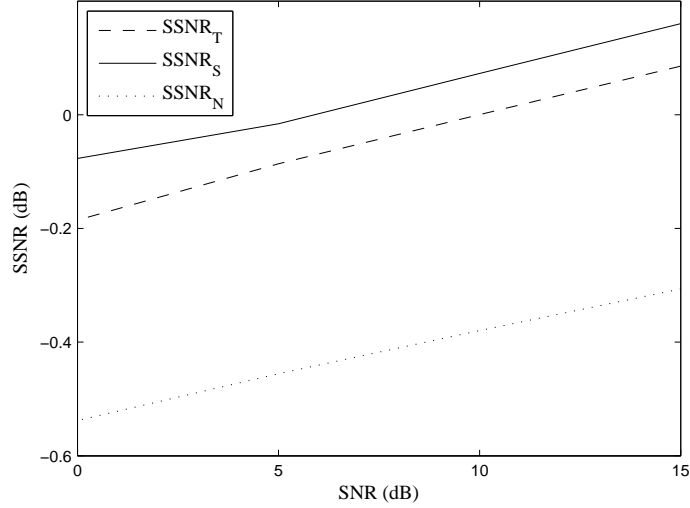


Figure 1.8: Segmental SNRs for noisy speech from the Noizeus database

$$\Delta_T = SSNR_T(x, \hat{x}) - SSNR_T(x, y) \quad (1.36)$$

$$\Delta_S = SSNR_S(x, \hat{x}) - SSNR_S(x, y) \quad (1.37)$$

$$\Delta_N = SSNR_N(x, \hat{x}) - SSNR_N(x, y) \quad (1.38)$$

Other quality metrics exist which may be more motivated for speech. [JM76] introduces various measures based on log-spectral distance, including the discretized COSH distance

$$d_{COSH}(x, \hat{x}) = \frac{1}{2N_{ch}} \sum_{k=1}^{N_{ch}} \left(\frac{A_k}{\hat{A}_k} + \frac{\hat{A}_k}{A_k} - 2 \right) \quad (1.39)$$

where N_{ch} is the number of channels used during spectral analysis. Figures 1.8 and 1.9 provide quantitative speech quality results for noisy speech from the Noizeus database [HL07].

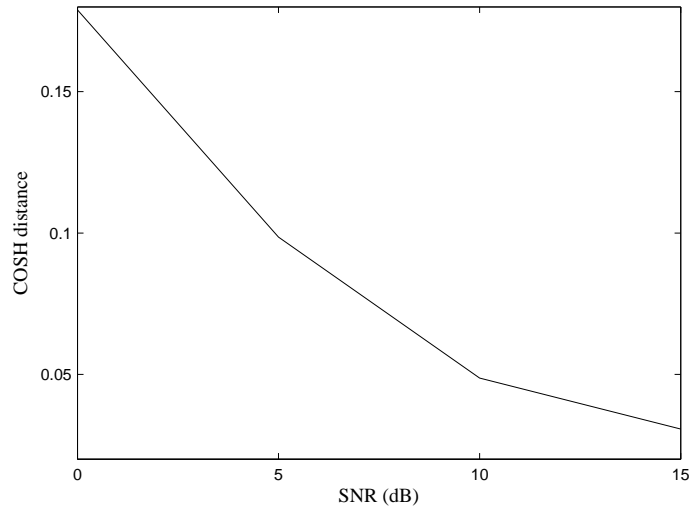


Figure 1.9: COSH distances [JM76] for noisy speech from the Noizeus database

1.3 Automatic Speech Recognition

Automatic speech recognition (ASR) serves as a widespread and useful application for human to computer interaction (HCI). While small-vocabulary tasks for controlled environments have achieved high accuracy rates, there currently exist many technological challenges within the field of ASR. These include handling large vocabulary tasks, speaker variability, and signal degradation due to environmental noise.

1.3.1 Feature Extraction

Feature extraction for ASR is designed to provide the recognizer with low-dimensionality features which maximize discriminative information. Whereas ASR front-ends can vary in design, certain components typically remain consistent.

The ASR feature extraction process is summarized below [RH93]:

- The input time-domain signal is windowed and transformed into its spectral representation. Most often, Fourier analysis is utilized, although wavelet analysis has

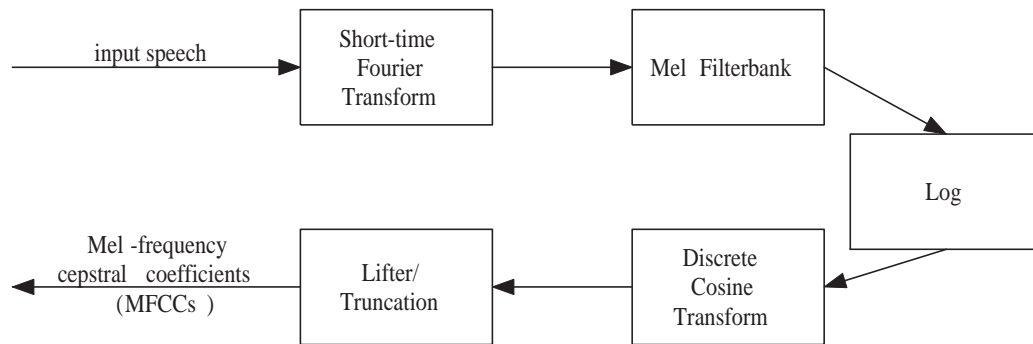


Figure 1.10: Overview of the MFCC computation for ASR front-end feature extraction

been proposed [GG01].

- The spectral representation is warped with respect to the frequency scale. Examples include the Mel scale and Bark scale [Qua01]. Frequency warping is most often based on human perception.
- The log operator is applied to feature vectors, effectively companding the input signal. Again, this step is motivated by the human auditory system.
- A further transform (typically the discrete cosine transform) is applied to decorrelate feature vector elements. The resulting domain is referred to as the cepstrum.
- A lifter and truncation are applied to reduce dimensionality, and to emphasize discriminative components [Pal99].

An illustrative example of the ASR front-end feature extraction process is provided in Figure 1.10.

1.3.2 Hidden Markov Models

A typical method for performing automatic recognition of speech is to model acoustic patterns with hidden Markov models (HMMs) [Rab89]. Although the theory of HMMs is applicable to a wide array of estimation problems, this section focuses on their use

in ASR. Define Ξ as an HMM comprised of discrete states $\{s_1, \dots, s_N\}$. Ξ can then be parameterized by the set of statistical measures $\{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$. Here, steady-state probabilities are provided by

$$\boldsymbol{\pi} = [\pi_1, \dots, \pi_N] \quad (1.40)$$

where

$$\pi_i = P(s_i(t)) \quad (1.41)$$

Transitional statistics are provided by the matrix \mathbf{A}

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \quad (1.42)$$

where

$$a_{ij} = P(s_j(t) | s_i(t-1)) \quad (1.43)$$

Let $\mathbf{O}_{1:T}$ be an observation sequence $\{\mathbf{o}_1, \dots, \mathbf{o}_T\}$. Observation statistics are then given by the set of probability distribution functions (pdfs) $\mathbf{B}=\{b_1(\mathbf{o}_t), \dots, b_N(\mathbf{o}_t)\}$, where

$$b_i(\mathbf{o}_t) = p(\mathbf{o}_t | s_i(t)) \quad (1.44)$$

Furthermore, let $S_{1:T}$ denote the sequence of occupied hidden states corresponding to time indices 1 through T .

In the context of ASR, HMMs are used to model acoustic patterns, such as words or phonemes. It follows that individual states model specific acoustic correlates which

comprise the global pattern. Observation pdfs convey statistical information regarding the likelihood of an acoustic observation given the occupancy of a certain hidden state.

There exist three fundamental problems associated with the use of HMMs for estimation ([Rab89], [RH93]):

1. Determining the probability of a sequence of observations given a parameterized HMM, $p(\mathbf{O}|\Xi)$
2. Determining the maximum a posteriori sequence of hidden states given an observation sequence and parameterized HMM, $S_{1:T}^o = \arg \max_{S_{1:T}} p(S_{1:T}|\mathbf{O}_{1:T}, \Xi)$
3. Estimating optimal HMM parameters given an observation sequence, $\Xi^o = \arg \max_{\Xi} p(\Xi|\mathbf{O}_{1:t})$

In the context of ASR, Problem 3 corresponds to the training process for acoustic models, and is widely discussed in literature, such as [Rab89] and [RH93]. Problems 1 and 2 pertain to the recognition process, and are summarized in this section.

Define the *forward variable* as the likelihood of $s_i(t)$ along with a sequence of observations up until and including time index t , given a parameterized HMM. The forward variable is determined recursively

$$\alpha_i(t) = P(\mathbf{O}_{1:t}, s_i(t) | \Xi) = \left[\sum_{j=1}^N \alpha_j(t-1) a_{ji} \right] b_i(\mathbf{o}_t) \quad (1.45)$$

Using Bayes' rule, the probability of a given observation sequence can be expressed as:

$$P(\mathbf{O}_{1:t}|\Xi) = \sum_{j=1}^N P(\mathbf{O}_{1:t}, s_j(t) | \Xi) = \sum_{j=1}^N \alpha_j(t) \quad (1.46)$$

which represents the solution to problem 1.

The solution to problem 2 requires state-specific occupancy probabilities at arbitrary time indices, given a full set of observations and given a parameterized HMM:

$$\gamma_i(t) = P(s_i(t) | \mathbf{O}_{1:T}, \Xi) \quad (1.47)$$

When future observations are accessible, the *backward variable* is defined as:

$$\beta_i(t) = P(\mathbf{O}_{t+1:T} | s_i(t), \Xi) = \sum_{j=1}^N \beta_j(t+1) a_{ij} b_j(\mathbf{o}_{t+1}) \quad (1.48)$$

Applying Bayes' rule leads to:

$$\gamma_i(t) = P(s_i(t) | \mathbf{O}_{1:T}, \Xi) = \frac{P(\mathbf{O}_{1:t}, s_i(t) | \Xi) P(\mathbf{O}_{t+1:T} | s_i(t), \Xi)}{P(\mathbf{O}_{1:T} | \Xi)} = \frac{\alpha_i(t) \beta_i(t)}{\sum_{j=1}^N \alpha_j(t) \beta_j(t)} \quad (1.49)$$

During automatic speech recognition, the maximum a posteriori state sequence is hypothesized:

$$s_i^o(t) = s_i(t) \quad \text{such that } i = \arg \max_j \gamma_j(t) \quad (1.50)$$

Although hidden Markov models are utilized for all ASR experimentation in this dissertation, HMMs are applied for various other tasks such as feature estimation (Chapters 4,6.1,6), and feature prediction (Chapter 8). In these applications, minimum mean-square error HMM estimation may be more suitable than the MAP solution from Eq. 1.50. If comprising states can be mapped to corresponding scalar values by the function \mathfrak{F} , the MMSE estimate is expressed as

$$\mathfrak{F}\{s_i^o(t)\} = \sum_{j=1}^N \gamma_j(t) \mathfrak{F}\{s_j\} \quad (1.51)$$

State mapping functions are present in problems involving scalar quantization, for example, where HMMs are used to infer the hidden quantizer centroid. In such a framework, the state mapping function outputs the scalar value corresponding to the input centroid.

1.3.3 Noise Robustness in ASR

The presence of background acoustic noise is well known to degrade ASR performance ([Ace93], [Gon95]). A wide variety of approaches exist for improving the robustness of ASR systems to corruptive noise, such as signal enhancement [Loi07], noise robust feature design ([ZA03], [Her90]), back-end weighted recognition ([CBA03], [IH07]), [RS05]), and multi-modal recognition ([Haz06], [BA08b]). Each of the previous approaches has been widely studied in literature. More recently, missing feature (MF) compensation has gained attention as an alternative approach to noise robust ASR.

1.3.4 Missing Feature Approaches to Noise Robust ASR

Missing feature algorithms have been shown successful in adverse acoustic environments [RS05]. Furthermore, MF approaches are argued in [CGJ01] to be inspired by the human auditory system. Missing feature algorithms for robust speech recognition can be grouped into two main approaches: back-end marginalization and front-end spectral reconstruction. Figures 1.11 and 1.12 provide overviews of front-end and back-end missing feature ASR systems, respectively. Each approach requires mask estimation to determine the reliability of individual time-frequency components. Various types of masks exist: soft masks provide soft reliabilities, whereas hard masks provide binary measures. Additionally, several studies involving missing feature approaches include results based on oracle masks [RS05]. Oracle masks assume exact knowledge of the clean version of the input

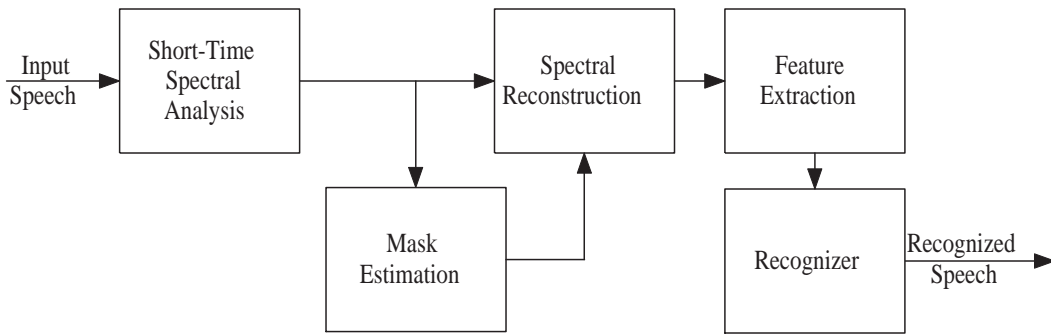


Figure 1.11: General overview of a front-end missing feature ASR system

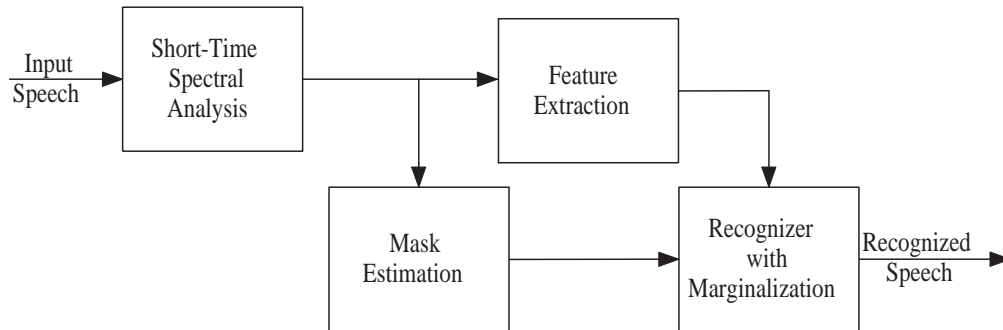


Figure 1.12: General overview of a back-end missing feature ASR system

speech signal, and are determined via time-frequency specific SNR comparisons against predetermined thresholds. Oracle masks can therefore be used to assess the success of spectral reconstruction or mask estimation techniques by providing upper performance bounds in each case.

In [VGC99], the authors present an algorithm for binary mask estimation based on spectral subtraction. Methods for soft-decision mask estimation include [GBC01], which uses a tunable sigmoid function to map SNR-related measures to the range $[0, 1]$. In [KS06] and [SRS04], Bayesian classifiers are trained to determine the empirical distribution of informative features in various noise types and levels. The mask estimation technique in [Sh04] utilizes phonetic class-dependent vector quantizer codebooks.

MF marginalization utilizes information regarding the reliability of spectral features in the back-end recognizer by deemphasizing posterior observation probabilities corre-

sponding to unreliable features ([BJC00], [CMG97], [CGJ01], [VGC99]). Similar MF-based techniques have been applied to channel mitigation for distributed speech recognition (DSR) ([IH06], [BA02]). Assume an ASR system is designed to perform recognition based on the spectral vector \mathbf{a}

$$\mathbf{a} = [A_1, \dots, A_M]^T \quad (1.52)$$

where M is the dimensionality of \mathbf{a} . Suppose \mathbf{a} is grouped into subsets of reliable and unreliable components, \mathbf{a}_r and \mathbf{a}_u , respectively. MF marginalization techniques adapt HMM observation statistics to compensate for unreliable components. Assuming uncorrelated observations

$$b_i(\mathbf{a}) = \prod_{A_j \in \mathbf{a}_r} b_i(A_j) \prod_{A_k \in \mathbf{a}_u} \int_{\epsilon_l}^{\epsilon_h} b_i(A_k) dA_k \quad (1.53)$$

where ϵ_h and ϵ_l denote upper and lower integration bounds, respectively. If unreliable features are completely ignored, then $\epsilon_h = \infty$ and $\epsilon_l = -\infty$, and the integral in Eq. 1.53 reduces to unity. However, soft information regarding the uncertainty of $\mathbf{a}_u(n)$ can be exploited by adapting integration bounds [CGJ01]. Although shown successful, back-end MF techniques generally involve extensive computation due to the integral terms present in Eq. 1.53.

Front-end spectral reconstruction techniques estimate unreliable spectral features in the front end so that recognition is based on estimated spectral information. The inference of degraded data is generally based on an underlying signal model. Additionally, if degraded observations are accessible, such observations can be exploited during reconstruction. In [Ram00] and [RSS04], reconstruction is performed by determining a MMSE-like estimate, given a trained Gaussian mixture model (GMM). The algorithm builds upon determining the bounded MAP estimate of unreliable components for each GMM cluster k

$$\hat{\mathbf{a}}_u^k = \arg \max_{\mathbf{a}_u} p(\mathbf{a}_u, \mathbf{a}_u \leq \mathbf{r}_u | k, \mathbf{r}_r) \quad (1.54)$$

where \mathbf{r}_r and \mathbf{r}_u are the subsets of reliable and unreliable observations, respectively. The unreliable components can then be estimated as

$$\mathbf{a}_u = \sum_k p(k | \mathbf{r}_r, \mathbf{a}_u \leq \mathbf{r}_u) \mathbf{a}_u^k \quad (1.55)$$

Computation of the posterior probability in Eq. 1.55 requires the following distribution

$$\begin{aligned} p(\mathbf{r}_r, \mathbf{a}_u \leq \mathbf{r}_u | k) &= \prod_{A_i \in \mathbf{a}_r} \frac{1}{\sqrt{2\pi\sigma_{k,i}^2}} \exp\left(-\frac{(A_i - \mu_{k,i})^2}{2\sigma_{k,i}^2}\right) \\ &\times \prod_{A_j \in \mathbf{a}_u} \int_{-\infty}^{R_j} \frac{1}{\sqrt{2\pi\sigma_{k,j}^2}} \exp\left(-\frac{(A_j - \mu_{k,j})^2}{2\sigma_{k,j}^2}\right) dA_j \end{aligned} \quad (1.56)$$

where $\mu_{k,i}$ and $\sigma_{k,i}^2$ are the mean and variance of the i^{th} component of the k^{th} GMM mixture. Note that Eq. 1.56 includes several irreducible integrals, which may prove computationally expensive. In [KH09], the GMM-based method was extended to include certain aspects of inter-frame correlation.

An issue with previous MF studies is the computational complexity. Specifically, [RSS04] reported single sentence processing times of ≈ 40 seconds and ≈ 10 seconds for back-end marginalization (Eq. 1.53) and front-end cluster-based reconstruction (Eq. 1.55), respectively.

1.4 Organization of Dissertation

This dissertation is organized as follows. Part I focuses on short-time spectral amplitude estimation for single-channel speech enhancement. Chapter 2 proposes unified frameworks for determining STSA estimators. Specifically, separate approaches are provided for the use of generalized Gamma-distributed priors, and for assuming phase equivalence of speech and noise. Chapter 3 explores the role of temporal correlation in STSA estimation. Finally, in Chapter 4, an algorithm for determining improved speech presence probabilities (SPPs) utilizing HMM-based inference, is proposed.

Part II explores front-end missing feature approaches to noise robust ASR. In Chapter 6.1, a statistical approach to Mel-domain mask estimation is discussed, which provides soft probabilistic metrics corresponding to individual spectral components. Next, two alternative MF spectral reconstruction solutions are introduced. Specifically, Chapter 6 proposes an HMM-based method of inferring missing data, whereas Chapter 7 explores the role of signal sparsity in spectral reconstruction. Finally, missing feature theory is extended to the task of packet loss concealment in Chapter 8. The dissertation is concluded by a summary and discussion of future work, in Chapter 9.

Part I

**Short-Time Spectral Amplitude
Estimation for Single-Channel Speech
Enhancement**

CHAPTER 2

Unified Frameworks for Deriving Short-Time Spectral Amplitude Estimators

This chapter proposes frameworks for deriving STSA estimators. Section 2.1 explores the use of GGD speech priors. Section 2.2 applies the assumption of phase equivalence of speech and noise. In each case, solutions are derived as functions of distribution shape parameters. In many cases, it is shown that particular solutions reduce to well-known traditional estimators.

2.1 A Unified Framework for STSA Estimation Using Generalized Gamma Distributions

This section presents a framework for determining MAP and MMSE solutions assuming generalized Gamma-distributed (GDD) speech priors. Proposed estimators are shown to reduce to well-known solutions for particular shape parameters. Furthermore, a novel speech presence filter is presented. Solutions are expressed in general form, as functions of GGD shape parameters, providing flexibility in modeling *a priori* speech STSA distributions.

To fit GGD distributions to the statistical distribution of speech, maximum likelihood estimation of shape parameters is applied. It is shown that empirically determined shape parameters exhibit a strong dependency on frequency channel. Motivated by this observation, channel-specific shape parameters are used during STSA estimation.

2.1.1 MAP Estimation

The MAP solution to STSA estimation is discussed in Section 1.2.4. Using Eq. 1.14 and the GGD speech prior from Eq. A.9, the general MAP estimator is given as the solution to

$$A_k^2 - \frac{\sigma_n^2(k)}{2} \frac{\partial}{\partial A_k} \log I_0 \left(\frac{2A_k R_k}{\sigma_n^2(k)} \right) A_k - \frac{(\zeta\nu - 1) \sigma_n^2(k)}{2} + \frac{\beta\zeta\sigma_n^2(k)}{2} A_k^\zeta = 0 \quad (2.1)$$

If the large-value approximation of $I_0(z)$ from Eq. A.20 is applied, Eq. 2.1 reduces to

$$A_k^2 - R_k A_k - \frac{(\zeta\nu - 3/2) \sigma_n^2(k)}{2} + \frac{\beta\zeta\sigma_n^2(k)}{2} A_k^\zeta = 0 \quad (2.2)$$

For shape parameter values of $\zeta=1$ or $\zeta=2$, the expression in Eq. 2.2 can be solved in closed-form using the quadratic equation

$$G_{MAP}(\xi_k, \gamma_k) |_{\zeta=2} = \frac{1}{2} \left(\frac{\xi_k}{\nu + \xi_k} \right) + \frac{1}{2} \sqrt{\left(\frac{\xi_k}{\nu + \xi_k} \right)^2 + \frac{2\xi_k(2\nu - 3/2)}{\gamma_k(\nu + \xi_k)}} \quad (2.3)$$

$$G_{MAP}(\xi_k, \gamma_k) |_{\zeta=1} = \frac{1}{2} \left(1 - \frac{\sqrt{\nu(\nu+1)}}{2\sqrt{\xi_k\gamma_k}} \right) + \frac{1}{2} \sqrt{\left(1 - \frac{\sqrt{\nu(\nu+1)}}{2\sqrt{\xi_k\gamma_k}} \right)^2 + \frac{2(\nu - 3/2)}{\gamma_k}} \quad (2.4)$$

Note that for $(\zeta=2, \nu=1)$, which corresponds to Rayleigh speech spectral amplitudes, the MAP estimator in Eq. 2.3 reduces to the solution presented by Wolfe and Godsil in [WG03]. Also, for $(\beta=1, \nu=2)$, which corresponds approximately to Laplacian speech spectral coefficients, the MAP estimator in Eq. 2.4 reduces to the solution presented by Lotter and Vary in [LV03] and [LV05].

2.1.2 MMSE Estimation

From Section 1.2.3, the minimum mean-square error solution to STSA estimation is given by

$$G_{MMSE}(\xi_k, \gamma_k) = \frac{1}{R_k} \frac{\int_0^\infty A_k p(Y_k|A_k) p(A_k) dA_k}{\int_0^\infty p(Y_k|A_k) p(A_k) dA_k} \quad (2.5)$$

Substitution of Eq. 1.14 and the GGD speech prior from Eq. A.9 yields

$$G_{MMSE}(\xi_k, \gamma_k) = \frac{1}{R_k} \frac{\int_0^\infty A^{\zeta\nu} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + R_k^2}{\sigma_n^2(k)} - \beta A_k^\zeta\right) \partial A_k}{\int_0^\infty A^{\zeta\nu-1} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + R_k^2}{\sigma_n^2(k)} - \beta A_k^\zeta\right) \partial A_k} \quad (2.6)$$

For the case when $\zeta=2$, Eq. 2.6 can be expressed as

$$G_{MMSE}(\xi_k, \gamma_k) |_{\zeta=2} = \frac{1}{R_k} \frac{\int_0^\infty \frac{2A^{\zeta\nu}}{\vartheta_k \sigma_n(k)} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + \alpha_k^2 R_k^2}{\vartheta_k \sigma_n^2(k)}\right) \partial A_k}{\int_0^\infty \frac{2A^{\zeta\nu-1}}{\vartheta_k \sigma_n(k)} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + \alpha_k^2 R_k^2}{\vartheta_k \sigma_n^2(k)}\right) \partial A_k} \quad (2.7)$$

which can be recognized as the ratio of moments of a Rician distribution with parameters $\vartheta_k \sigma_n^2(k)/2$ and $\vartheta_k R_k$, where

$$\vartheta_k = \frac{\xi_k}{\nu + \xi_k} \quad (2.8)$$

Applying Eq. A.4 to Eq. 2.7 yields

$$G_{MMSE}(\xi_k, \gamma_k) |_{\zeta=2} = \sqrt{\frac{\vartheta_k}{\gamma_k}} \frac{\Gamma(\nu + 1/2)}{\Gamma(\nu)} \frac{{}_1F_1(-\nu + 1/2; 1; -\vartheta_k \gamma_k)}{{}_1F_1(-\nu + 1; 1; -\vartheta_k \gamma_k)} \quad (2.9)$$

A similar solution is proposed in [EHH07]. Note that for $(\zeta=2, \nu=1)$, which corresponds to Rayleigh speech spectral amplitudes, the MMSE solution reduces to that proposed by Ephraim and Malah in [EM84].

If instead $\zeta=1$, Eq. 2.6 becomes

$$G_{MMSE}(\xi_k, \gamma_k) |_{\zeta=1} = \frac{\int_0^\infty \frac{2A^\nu}{\sigma_n(k)} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + R_k^2}{\sigma_n^2(k)} - \frac{\sqrt{\nu(\nu+1)} A_k}{\sigma_x(k)}\right) \partial A_k}{\int_0^\infty \frac{2A^{\nu-1}}{\sigma_n(k)} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + R_k^2}{\sigma_n^2(k)} - \frac{\sqrt{\nu(\nu+1)} A_k}{\sigma_x(k)}\right) \partial A_k} \quad (2.10)$$

Using the large-value approximation of I_0 from Eq. A.20 leads to

$$G_{MMSE}(\xi_k, \gamma_k) |_{\zeta=1} = \frac{(\nu - 1/2) D_{-\nu-1/2}\left(\sqrt{\frac{\nu(\nu+1)}{2\xi_k}} - \sqrt{2\gamma_k}\right)}{\sqrt{2\gamma_k} D_{-\nu+1/2}\left(\sqrt{\frac{\nu(\nu+1)}{2\xi_k}} - \sqrt{2\gamma_k}\right)} \quad (2.11)$$

where D_ν is the parabolic cylinder function of order ν (see Appendix A.2.5). Note that the solution of Eq. 2.11 was previously presented in [EHH07].

Figure 2.1 illustrates gain curves for MAP (Eq. 2.3) and MMSE (Eq. 2.10) estimators corresponding to *a priori* speech distributions with $(\zeta=2, \nu=1)$. It can be observed that the

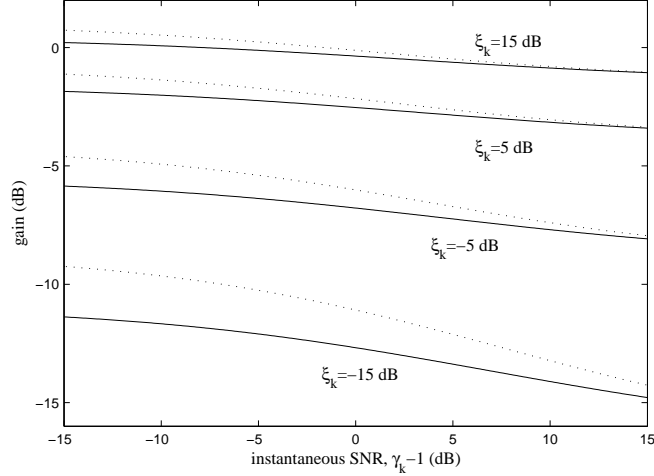


Figure 2.1: Gain curves for MAP (solid lines) and MMSE (dotted lines) estimators for generalized Gamma distributed speech with $(\zeta=2, \nu=1)$

MAP and MMSE solutions exhibit similar behavior, with MMSE attenuation decreasing more rapidly for decreasing γ_k .

Figure 2.2 illustrates gain curves for MAP (Eq. 2.4) and MMSE (Eq. 2.11) estimators corresponding to *a priori* speech distributions with $(\zeta=1, \nu=2)$. It can be observed that MAP and MMSE solutions exhibit very different behavior. Specifically, the MAP solution applies significantly more attenuation, especially at low values of ξ_k . This is due to the prominent tail of the GGD with $(\zeta=1, \nu=2)$, which is accounted for during MMSE estimation, but not during MAP estimation.

2.1.3 Estimation Under Speech Presence Uncertainty

Following Section 1.2.5, soft-decision speech presence probability filters can be derived for the case of generalized Gamma distributed speech STSAs. The conditional distribution given the hypothesis H_0 can be adapted from the marginal distribution of noise

$$p(Y_k | q_k = H_0) = \frac{1}{\pi \sigma_{N,k}^2} \exp\left(-\frac{R_k^2}{\sigma_{N,k}^2}\right) \quad (2.12)$$

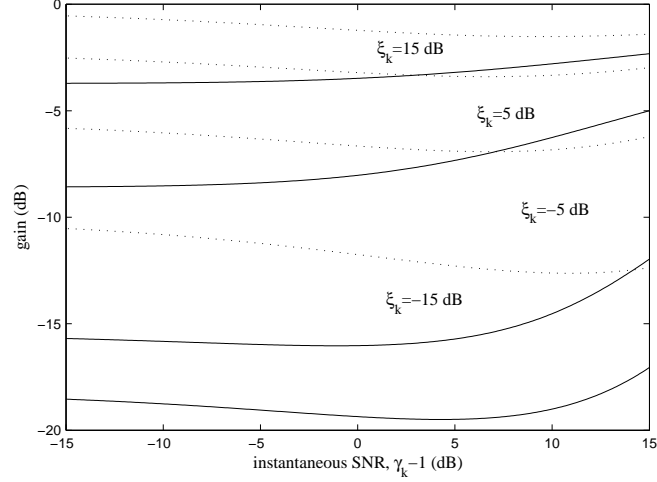


Figure 2.2: Gain curves for MAP (solid lines) and MMSE (dotted lines) estimators for generalized Gamma distributed speech with ($\zeta=1, \nu=2$)

The conditional distribution given an active speech state is determined via marginalization

$$p(Y_k | q_k = H_1) = \int_0^\infty p(Y_k | A_k) p(A_k) \partial A_k \quad (2.13)$$

Substitution of Eq. 1.14 and Eq. A.9 yields

$$p(Y_k | q_k = H_1) = \frac{\beta^\nu \zeta}{\pi \sigma_n^2(k) \Gamma(\nu)} \int_0^\infty A^{\zeta \nu - 1} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2 + R_k^2}{\sigma_n^2(k)} - \beta A_k^\zeta\right) \partial A_k \quad (2.14)$$

Substitution of 2.14 into Eq. 1.32 leads to

$$\Lambda(\xi_k, \gamma_k) = \frac{\eta_k \beta^\nu \zeta}{\Gamma(\nu)} \int_0^\infty A^{\zeta \nu - 1} I_0\left(\frac{2A_k R_k}{\sigma_n^2(k)}\right) \exp\left(-\frac{A_k^2}{\sigma_n^2(k)} - \beta A_k^\zeta\right) \partial A_k \quad (2.15)$$

For the case when $\zeta=2$, the expression for the GLR reduces to

$$\begin{aligned} \Lambda(\xi_k, \gamma_k) |_{\zeta=2} &= \frac{\eta_k \vartheta_k \sigma_n^2(k)}{\Gamma(\nu)} \left(\frac{\nu}{\sigma_x^2(k)} \right)^\nu \exp(\vartheta_k \gamma_k) \\ &\times \int_0^\infty \frac{2A^{2\nu-1}}{\vartheta_k \sigma_n^2(k)} I_0 \left(\frac{2A_k R_k}{\sigma_n^2(k)} \right) \exp \left(-\frac{A_k^2 + \alpha_k^2 R_k^2}{\vartheta_k \sigma_n^2(k)} \right) \partial A_k \end{aligned} \quad (2.16)$$

Note that the integral in Eq. 2.16 represents the $(2\nu - 2)^{th}$ moment of a Rician distribution with parameters $\vartheta_k R_k$ and $\lambda_k \sigma_n^2(k) / 2$, leading to

$$\Lambda(\xi_k, \gamma_k) |_{\zeta=2} = \eta_k \left(\frac{\nu}{\nu + \xi_k} \right)^\nu \exp(\vartheta_k \gamma_k) {}_1F_1(-\nu + 1; 1; -\vartheta_k \gamma_k) \quad (2.17)$$

For the case of $\nu=1$, corresponding to Gaussian spectral coefficients, the GLR is equivalent to the that proposed by Ephraim and Malah in [EM84] (Eq. 1.34).

For the case when $\zeta=1$, the GLR expression from Eq. 2.15 reduces to

$$\begin{aligned} \Lambda(\xi_k, \gamma_k) |_{\zeta=1} &= \frac{\eta_k}{\Gamma(\nu)} \left(\frac{\nu(\nu+1)}{\sigma_x^2(k)} \right)^{\nu/2} \\ &\times \int_0^\infty A_k^{\zeta\nu-1} I_0 \left(\frac{2A_k R_k}{\sigma_n^2(k)} \right) \exp \left(-\frac{A_k^2}{\sigma_n^2(k)} - \frac{\sqrt{\nu(\nu+1)} A_k}{\sigma(k)} \right) \partial A_k \end{aligned} \quad (2.18)$$

Applying the large-value approximation of I_0 from Eq. A.20 leads to

$$\begin{aligned} \Lambda(\xi_k, \gamma_k) |_{\zeta=1} &= \frac{\eta_k}{\Gamma(\nu)} \left(\frac{\nu(\nu+1)}{\sigma_x^2(k)} \right)^{\nu/2} \left(\frac{\sigma_n^2(k)}{2\sqrt{2\pi} R_k} \right) \\ &\times \int_0^\infty A_k^{\zeta\nu-1} \exp \left(-\frac{A_k^2}{\sigma_n^2(k)} - \frac{\sqrt{\nu(\nu+1)} A_k}{\sigma_x(k)} + \frac{2R_k A_k}{\sigma_n^2(k)} \right) \partial A_k \end{aligned} \quad (2.19)$$

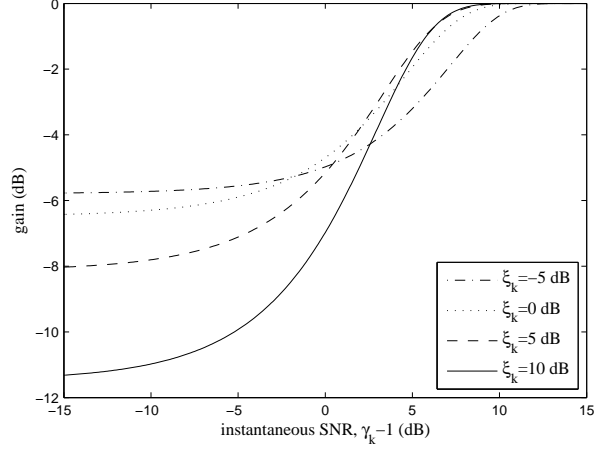


Figure 2.3: Gain curves for the proposed soft-decision speech presence probability filter with ($\zeta=1, \nu=2$)

which can be expressed using the parabolic cylinder function [AS65] as

$$\Lambda(\xi_k, \gamma_k) |_{\zeta=1} = \eta_k \frac{\Gamma(\nu - 1/2)}{\Gamma(\nu)} \left(\sqrt{\frac{\nu(\nu+1)}{2\xi_k}} \right)^\nu \left(\frac{1}{\sqrt{2\pi}\gamma_k^{1/4}} \right) \exp\left(\frac{\mu_k^2}{4}\right) D_{-\nu+1/2}(\mu_k) \quad (2.20)$$

where

$$\mu_k = \sqrt{\frac{\nu(\nu+1)}{2\xi_k}} - \sqrt{2\gamma_k} \quad (2.21)$$

Figure 2.3 illustrates gain curves for the soft-decision speech presence probability filter for ($\beta=1, \nu=2$). Note the similarities in behavior between SPP filters corresponding to shape parameters ($\zeta=1, \nu=2$), provided in Figure 2.3, and ($\zeta=2, \nu=1$), provided in Figure 1.7.

2.1.4 GGD Shape Parameter Estimation

Recent work involving super-Gaussian priors for speech spectral amplitudes are motivated by studies of empirical distributions obtained from training data ([Mar05], [LV03]). In such work, empirical pdfs hint at an underlying distribution which exhibits a narrow peak and well-defined tail. As discussed earlier, use of GGDs allows flexibility in capturing the statistical behavior of speech. However, the use of GGDs requires estimation of shape parameters ν and β , given $\zeta \in \{1, 2\}$. This section presents a maximum likelihood technique for GGD parameter estimation, similar to [CW69].

Let the likelihood function $\mathcal{L}(\beta, \nu | \mathbf{A})$ represent the probability of hypothesis shape parameters, conditioned on N data samples

$$\mathbf{A} = \{A_k(1), \dots, A_k(N)\} \quad (2.22)$$

Assuming samples are independent and identically distributed according to the GGD distribution from Eq. A.9, the likelihood is expressed as

$$\mathcal{L}(\beta, \nu | \mathbf{A}) = \prod_{i=1}^N \frac{\zeta \beta^\nu}{\Gamma(\nu)} A_k(i) \exp\left(-\beta A_k^\zeta(i)\right) \quad (2.23)$$

Applying the log function results in

$$\begin{aligned} \log \mathcal{L}(\beta, \nu | \mathbf{A}) = & \quad (2.24) \\ N \log \zeta + N \nu \log \beta - N \log \Gamma(\nu) + (\zeta \nu - 1) \sum_{i=1}^N \log A_k(i) - \beta \sum_{i=1}^N A_k^\zeta(i) \end{aligned}$$

Differentiation of Eq. 2.24 with respect to β , and equating to zero, leads to an expression for the estimated parameter β

$$\beta = \frac{N\nu}{\sum_{i=1}^N A_k^\zeta(i)} \quad (2.25)$$

Differentiating Eq. 2.24 with respect to ν , equating to zero, and substituting Eq. 2.25 results in

$$\Psi(\nu) - \log \nu + \log \frac{\sum_{i=1}^N A_k^\zeta(i)}{N} - \frac{\zeta}{N} \sum_{i=1}^N \log A_k(i) = 0 \quad (2.26)$$

where Ψ denotes the Digamma function (see Appendix A.2.2). No closed-form solution exists for Eq. 2.26. The bisection method is therefore applied to find the approximation ν . Note that complexity involved with the iterative bisection method is not an issue, since parameter fitting is generally performed off-line. Figure 2.4 illustrates estimated shape parameters ν as a function of frequency channel, for the $\zeta=1$ and $\zeta=2$ case. Note that the large ν values for very high and very low channels is a result of attenuation applied at these frequencies by the MIRS filter (see Section 1.2.6). Motivated by the observed dependence of ν on frequency, channel-specific (CS) shape parameters are used during STSA estimation.

2.1.5 Experimental Results

To assess the performance of the proposed speech enhancement methods, the Noizeus database is used, and results are averaged across all 30 utterances, and all 8 noise types. Proposed STSA estimators are embedded into code from [Coh], which applies a gain floor of -18 dB and online noise estimation according to [Coh02]. The DD approach for estimating the *a priori* SNR is utilized due to its simple yet effective implementation. Table 2.1 provides speech enhancement results for MAP and MMSE STSA estimators. The table includes results for shape parameters invariant of channel, set to well-known combinations ($\zeta=1, \nu=2$) and ($\zeta=2, \nu=1$), corresponding to Rayleigh and Laplacian distri-

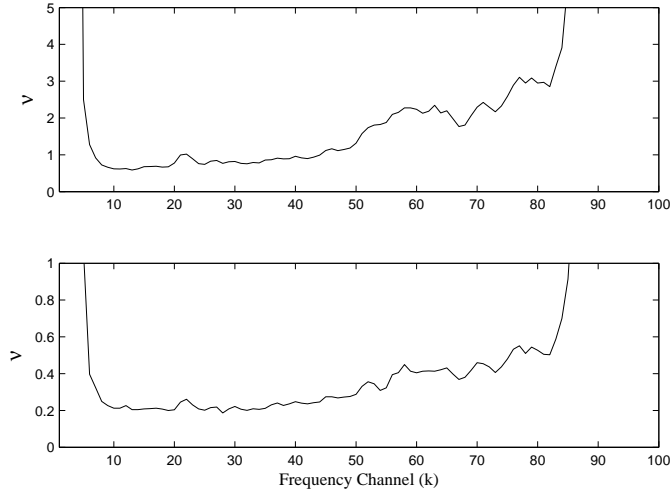


Figure 2.4: Estimated shape parameter ν as a function of frequency channel, for the $\zeta=1$ case (top panel) and $\zeta=2$ case (bottom panel)

butions. Results are also included for the proposed channel-specific (CS) method of shape parameter estimation from Section 2.1.4. It should be noted that channel-specific shape parameters are bounded for the sake of numerical stability, such that $\nu \in [0.5, 3]$ for $\zeta=1$, and $\nu \in [0.5, 1]$ for $\zeta=2$. Finally, results are included for combined STSA estimators exploiting speech presence uncertainty.

In Table 2.1, MMSE and MAP estimators derived within the proposed framework show promising enhancement ability. Additionally, proposed channel-specific ML parameter fitting is shown to improve MMSE estimators, particularly in poor conditions. Finally, the use of novel soft-decision speech uncertainty filters further improves noise suppression during enhancement, in terms of SSNR-related measures. Note that SSNR-related metrics are defined in Section 1.2.6.

Table 2.2 provides COSH distance measures for speech enhancement. Again, it can be concluded that the use of CS shape parameters provides improved signal quality for both MMSE and MAP estimation. Additionally, the use of SPP filters provides further improvement.

Table 2.1: Segmental SNR scores for MAP and MMSE STSA estimators: Bold entries denote the best score for each metric at each noise condition.

Criterion	(ζ, ν)	10 dB			5 dB			0 dB		
		Δ	Δ_S	Δ_N	Δ	Δ_S	Δ_N	Δ	Δ_S	Δ_N
without speech presence probability filter										
MMSE	(2, 1)	4.7	3.1	8.4	5.4	4.0	8.6	6.1	4.9	8.6
	(2, CS)	5.2	3.3	9.7	6.0	4.3	9.9	6.7	5.3	9.8
	(1, 2)	5.9	3.4	11.8	6.9	4.6	12.2	7.8	5.9	12.1
	(1, CS)	6.1	3.4	12.7	7.1	4.7	13.1	8.1	6.0	13.0
MAP	(2, 1)	5.6	3.4	10.8	6.5	4.5	11.1	7.3	5.6	11.0
	(2, CS)	5.9	3.5	11.8	6.8	4.7	12.1	7.7	5.8	12.0
	(1, 2)	5.5	3.0	11.3	6.6	4.3	12.2	7.7	5.6	12.5
	(1, CS)	5.7	3.3	11.5	6.7	4.5	12.1	7.7	5.7	12.2
with speech presence probability filter										
MMSE	(2, 1)	5.3	3.3	10.0	6.1	4.4	10.2	6.9	5.4	10.2
	(2, CS)	5.5	3.4	10.5	6.3	4.5	10.7	7.1	5.5	10.6
	(1, 2)	6.1	3.4	12.7	7.1	4.7	13.1	8.1	6.0	13.0
	(1, CS)	6.2	3.4	13.1	7.2	4.7	13.6	8.2	6.0	13.5
MAP	(2, 1)	5.9	3.5	11.6	6.8	4.6	11.9	7.6	5.8	11.8
	(2, CS)	6.0	3.5	11.9	6.9	4.7	12.2	7.7	5.9	12.2
	(1, 2)	5.8	3.2	12.1	6.9	4.4	12.9	7.9	5.8	13.1
	(1, CS)	6.0	3.4	12.2	7.0	4.6	12.7	7.9	5.9	12.7

Informal listening tests are consistent with observations from quantitative analysis in Table 2.1. For reasonable acoustic conditions (10 dB), speech quality is significantly improved with the use of STSA estimators derived within the proposed framework. In poor acoustic conditions (5 and 0 dB), noise suppression is audible, although some musical noise is noticeable for highly non-stationary noise, such as *babble* or *restaurant*. Furthermore, it is perceptually noticeable that the use of the novel soft-decision speech uncertainty filter increases noise suppression during speech enhancement.

Table 2.2: COSH distortion measures for MAP and MMSE STSA estimators. Bold entries denote the best score for each metric at each noise condition.

Estimator	(ζ_n, ν_n)	15 dB	10 dB	5 dB	0 dB
without speech presence probability filter					
MMSE	(2, 1)	1.75	3.36	5.88	10.60
	(2, CS)	1.55	2.95	5.14	9.25
	(1, 2)	1.33	2.38	4.06	7.19
	(1, CS)	1.33	2.35	3.95	6.98
MAP	(2, 1)	1.45	2.68	4.55	8.10
	(2, CS)	1.40	2.47	4.16	7.33
	(1, 2)	1.47	2.57	4.16	7.12
	(1, CS)	1.45	2.58	4.21	7.29
with speech presence probability filter					
MMSE	(2, 1)	1.51	2.83	4.95	8.87
	(2, CS)	1.46	2.70	4.70	8.41
	(1, 2)	1.30	2.25	3.76	6.58
	(1, CS)	1.30	2.22	3.71	6.48
MAP	(2, 1)	1.39	2.48	4.24	7.49
	(2, CS)	1.39	2.42	4.09	7.19
	(1, 2)	1.38	2.41	3.92	6.72
	(1, CS)	1.37	2.40	3.97	6.88

2.2 A Unified Framework for STSA Estimation Assuming Phase Equivalence of Speech and Noise

2.2.1 Phase Equivalence and Spectral Subtraction

As early as [Bol79], the notion of phase equivalence in spectral magnitude estimation was studied by Boll, leading to the well-known magnitude spectral subtraction (MSS) solution. Since then, multiple variations of spectral subtraction, such as power spectral subtraction (PSS) [BSM79], and nonlinear spectral subtraction (NSS) [LB92], have been developed both for speech enhancement and noise robust automatic speech recognition (ASR).

In this study, the role of phase equivalence is explored for statistical approaches to STSA estimation. In [Bol79], spectral observations are interpreted as deterministic values, implying underlying spectral speech and noise components to be deterministic as well. In this study, however, a stochastic approach is applied, and spectral observations, and thus the underlying speech and noise components, are interpreted as random processes.

Applying the assumption of phase equivalence of speech and noise spectral components, i.e. $\alpha_k = \psi_k$, leads Eq. 1.3 to be expressed as

$$R_k = A_k + D_k \quad (2.27)$$

This assumption reduces the complexity required by the estimation process, as it effectively projects the spectral estimation problem onto a 1-dimensional subspace of the complex domain. As discussed in [LL08], the relationship of Eq. 2.27, which is implied by the assumption of phase equivalence, results in a speech estimate that is less than or equal to the actual clean speech. Specifically, the law of cosines leads to (see Appendix A.3)

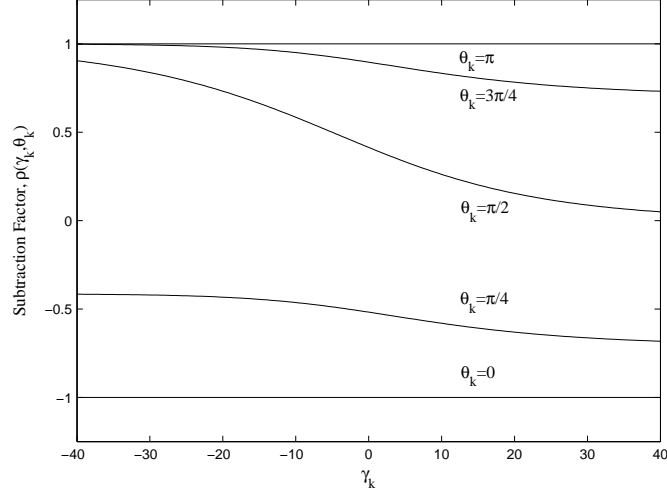


Figure 2.5: Subtraction factors, ρ_k , as a function of γ_k and θ_k

$$A_k = R_k - \rho_k(\gamma_k, \theta_k) D_k \quad (2.28)$$

where

$$\theta_k = \alpha_k + (\pi - \psi_k) \quad (2.29)$$

and

$$\rho_k(\gamma_k, \theta_k) = \sqrt{\gamma_k} - \sqrt{\gamma_k - \sin^2 \theta_k} - \cos \theta_k \quad (2.30)$$

Figure 2.5 illustrates the subtraction factor $\rho_k(\gamma_k, \theta_k)$ as a function of *a posteriori* SNR and relative angle θ_k . It can be observed that increasing levels of the *a posteriori* SNR lead to decreased attenuation.

Based on Figure 2.5, it can be concluded that spectral subtraction methods such as MSS [Bol79], which correspond to the $\theta_k = \pi$ case, generally result in over-attenuation. As previously mentioned, this study explores a stochastic approach to spectral subtraction,

and therefore proposed solutions may be expected to suffer from similar over-attenuation. To compensate for this, a subtraction factor, $\tilde{\rho}_k$, can be applied

$$R_k = A_k + \tilde{\rho}_k D_k, \text{ for } 0 \ll \rho_k \leq 1 \quad (2.31)$$

yielding modified definitions of the *a priori* and *a posteriori* SNR

$$\begin{aligned} \tilde{\xi}_k &= \frac{\sigma_x^2(k)}{\tilde{\rho}_k^2 \sigma_n^2(k)} \\ \tilde{\gamma}_k &= \frac{R_k^2}{\tilde{\rho}_k^2 \sigma_n^2(k)} \end{aligned} \quad (2.32)$$

The subtraction factor is chosen according to Eq. 2.5, evaluated at some relative angle. In this study, a value of $\theta_k=7\pi/8$ provided promising results, when tested on the Noizeus database [HL07]. Note that in Sections 2.2.2-2.2.4, STSA estimators are derived in terms of standard values γ_k and ξ_k for the sake of consistency with existing solutions. However, a subtraction factor can be applied by instead using SNR values from Eq. 2.32.

Studies in [LV05] and [Mar05] provide empirical *a priori* distributions for speech and noise. Non-monotonic, unimodal distributions are shown for both speech and noise spectral amplitudes, similar to Erlang-2 and Rayleigh random variables. It should be noted, however, that empirical studies may often be greatly dependent on data, and may not convey true statistics. Therefore, estimation solutions are derived as functions of GGD shape parameters when mathematically possible. In this way, specific solutions can be obtained by substituting those shape parameters corresponding to expected speech and noise prior distributions.

2.2.2 ML Estimation

In this section, ML short-time spectral estimators based on the aforementioned phase equivalence assumption are presented. ML estimation offers an efficient framework for inferring unknown parameters when *a priori* distributions of the target signal are not known. From Eq. 2.27, it can be concluded that the conditional probability of observed spectral components, given underlying clean components, is simply dependent on noise statistics

$$p(R_k|A_k) = p(D_k = R_k - A_k) = \frac{\zeta_n \beta_n^{\nu_n}}{\Gamma(\nu_n)} (R_k - A_n)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n}\right) \quad (2.33)$$

Using Eq. 1.17, the ML solution is generalized as

$$G_{ML}(\xi_k, \gamma_k) = 1 - \frac{1}{\sqrt{\gamma_k}} \left(\frac{\zeta_n \nu_n - 1}{\zeta_n \sqrt{\nu_n (\nu_n + 2 - \zeta_n)}} \right)^{1/\zeta_n}, \text{ for } \zeta_n \in \{1, 2\} \quad (2.34)$$

Analysis of the second derivative of Eq. 2.33 reveals that for $\zeta_n \nu_n < 1$, the solution of Eq. 2.34 does not exist, since the distribution $p(R_k, A_k)$ is monotonic, and thus includes no maximum. For $\zeta_n=1$ and $\nu_n=1$, the G_{ML} estimator reduces to unity.

It is interesting to note the similarity between the G_{ML} estimator and magnitude spectral subtraction presented by Boll in [Bol79]. Particularly, for $\zeta_n=2$ and for large values of ν_n , which correspond to deterministic values without uncertainty, the proposed ML estimator approaches that given in [Bol79].

$$G_{ML}(\xi_k, \gamma_k) |_{\zeta_n=2, \nu_n \rightarrow \infty} = \frac{\sqrt{\gamma_k} - 1}{\sqrt{\gamma_k}} \quad (2.35)$$

Figure 2.6 illustrates gain curves for the proposed ML STSA estimator for $\zeta_n=2$, and for

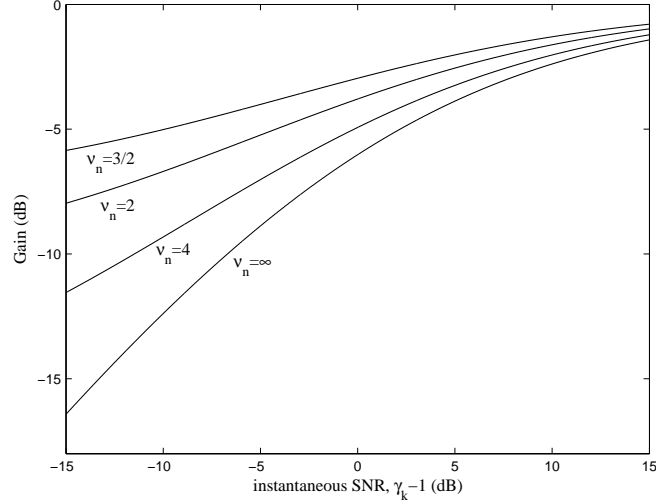


Figure 2.6: Gain curves for the G_{ML} STSA estimator, with $\zeta_n=2$, and for various values of ν_n . Note that for $\nu_n=\infty$, the proposed ML estimator is equivalent to the magnitude spectral subtraction solution from [Bol79].

various values of the GGD shape parameter ν_n .

2.2.3 MMSE Estimation

This section derives MMSE short-time spectral amplitude estimators assuming phase equivalence of speech and noise. By assuming phase equivalence of speech and noise components, i.e.

$$p(\alpha_k) = \delta(\alpha_k - \varphi_k), \quad (2.36)$$

where $\delta(\cdot)$ represents the Dirac delta function, the MMSE solution from Eq. 2.7 simplifies to

$$\hat{A}_k = \frac{\int_0^\infty A_k p(R_k|A_k) p(A_k) \partial A_k}{\int_0^\infty p(R_k|A_k) p(A_k) \partial A_k} \quad (2.37)$$

Using Eq. 2.33, and the fact that $A_k, D_k \in [0, R_k]$, Eq. 2.7 reduces to

$$\hat{A}_k = \frac{\int_0^{R_k} A_k p(D_k = R_k - A_k) p(A_k) \partial A_k}{\int_0^{R_k} p(D_k = R_k - A_k) p(A_k) \partial A_k} \quad (2.38)$$

Assuming generalized Gamma distributions for speech and noise spectral magnitudes, the solution of Equation 2.38 becomes

$$\hat{A}_k = \frac{\int_0^{R_k} A_k^{\zeta_x \nu_x} (R_k - A_k)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n} - \beta_x A_k^{\zeta_x}\right) \partial A_k}{\int_0^{R_k} A_k^{\zeta_x \nu_x - 1} (R_k - A_k)^{\zeta_n \nu_n - 1} \exp\left(-\beta_n (R_k - A_k)^{\zeta_n} - \beta_x A_k^{\zeta_x}\right) \partial A_k} \quad (2.39)$$

Equation 2.39 provides a general solution to MMSE spectral magnitude estimation assuming phase equivalence, and assuming priors from the generalized Gamma family. Particular solutions can be obtained by substituting specific shape parameters corresponding to desired speech and noise distributions.

Due to varying conclusions regarding the true statistical behavior of speech, assumptions of specific models for speech and noise are avoided. Instead, Equation 2.39 is provided as a function of GGD shape parameters. The following subsection derives MMSE solutions for shape parameters which may be considered realistic for speech and noise.

2.2.3.1 Assuming Gaussian Noise Priors ($\zeta_n=2, \nu_n=1/2$)

In this section, MMSE spectral magnitude estimation is discussed for Gaussian noise priors. Specifically, separate particular solutions are presented assuming Gaussian and exponential speech priors, referred to as *GGMMSE* and *GEMMSE*, respectively.

The GGMMSE solution, which utilizes Gaussian speech and noise priors, is derived from Equation 2.39

$$\hat{A}_k = \frac{\int_0^{R_k} A_k \exp(-\beta_n (R_k - A_k)^2 - \beta_x A_k^2) \partial A_k}{\int_0^{R_k} \exp(-\beta_n (R_k - A_k)^2 - \beta_x A_k^2) \partial A_k} \quad (2.40)$$

The integrals in Eq. 2.40 can be solved to reveal the GGMMSE gain function

$$G_{GGMMSE}(\xi_k, \gamma_k) = \phi_k - \sqrt{\frac{2\phi_k}{\pi\gamma_k}} \left(\frac{\exp\left(-\frac{\gamma_k}{2\xi_k(1+\xi_k)}\right) - \exp\left(-\frac{\gamma_k\phi_k}{2}\right)}{\operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2\xi_k(1+\xi_k)}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k\phi_k}{2}}\right)} \right) \quad (2.41)$$

where ϕ_k is the traditional Wiener filter (WF) [Wie49]

$$\phi_k = \frac{\xi_k}{1 + \xi_k}. \quad (2.42)$$

Additionally, $\operatorname{erf}(\cdot)$ represents the Gauss error function (see Eq. A.12). It is interesting to note that for extreme values of the *a posteriori* SNR, the GGMMSE estimator approximates the Wiener filter

$$G_{GGMMSE}(\xi_k, \gamma_k) \Big|_{|\gamma| \gg 1} \approx \phi_k. \quad (2.43)$$

More specifically, the GGMMSE gain function can be interpreted as the Wiener filter with an additive modification factor. To avoid a solution which includes irreducible functions, the Gauss error function is approximated by its truncated Taylor series expansion. The Taylor series expansion of the Gauss error function is given by Eq. A.14. Using the 1st-order approximation of the Gauss error function, the \hat{G}_{GGMMSE}^1 solution becomes

$$\hat{G}_{GGMMSE}^1(\xi_k, \gamma_k) = \phi \left[1 - \frac{1}{\gamma_k} \exp\left(-\frac{\gamma_k(1+\xi_k^2)}{4\xi_k(1+\xi_k)}\right) \sinh\left(\frac{\gamma_k(\xi_k-1)}{4\xi_k}\right) \right] \quad (2.44)$$

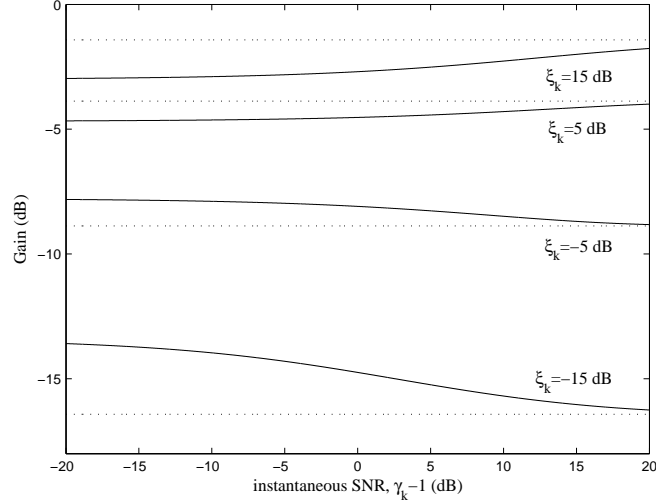


Figure 2.7: Gain curves for the \hat{G}_{GGMMSE}^1 STSA estimator (solid line) for various values of ξ_k : The Wiener filter [Wie49] (dotted line) is included for comparison.

Figure 2.7 illustrates gain curves for the \hat{G}_{GGMMSE}^1 estimator, for various *a priori* SNRs. The Wiener filter is included as a reference. As can be observed in Figure 2.7, the \hat{G}_{GGMMSE}^1 converges to the Wiener filter for large values of the *a posteriori* SNR. Additionally, for favorable acoustic conditions ($\xi_k=15,5$ dB), the \hat{G}_{GGMMSE}^1 estimator provides increasing attenuation as the *a posteriori* SNR decreases, which follows intuitively. For unfavorable conditions ($\xi_k=-5,-15$ dB), however, increased attenuation is applied for increasing *a posteriori* SNR. As discussed in [EM84], such behavior is the result of the estimator compromising between *a priori* information in the form of ξ_k , and new information introduced by γ_k .

Retaining the previous Gaussian model for the noise component, the GEMMSE solution can be adapted from Equation 2.39 by assuming exponential speech priors

$$\hat{A}_k = \frac{\int_0^{R_k} A_k \exp(-\beta_n (R_k - A_k)^2 - \beta_x A_k) \partial A_k}{\int_0^{R_k} \exp(-\beta_n (R_k - A_k)^2 - \beta_x A_k) \partial A_k} \quad (2.45)$$

The GEMMSE gain function can be derived from Equation 2.45 as

$$G_{GEMMSE}(\xi_k, \gamma_k) = 1 - \sqrt{\frac{2}{\xi_k \gamma_k}} - \sqrt{\frac{2}{\pi \gamma_k}} \left[\frac{\exp\left(-\frac{1}{\xi_k}\right) - \exp\left(-\frac{\gamma_k}{2} - \frac{1}{\xi_k} + \frac{\sqrt{2\gamma_k}}{\sqrt{\xi_k}}\right)}{\operatorname{erf}\left(\frac{1}{\sqrt{\xi_k}}\right) + \operatorname{erf}\left(\sqrt{\frac{\gamma_k}{2}} - \frac{1}{\sqrt{\xi_k}}\right)} \right] \quad (2.46)$$

As in the previous section, the GEMMSE includes irreducible Gauss error functions, which can be expressed as Taylor series expansions. Following steps from the previous section, the GEMMSE solution can be approximated as

$$\hat{G}_{GEMMSE}^1(\xi_k, \gamma_k) = \quad (2.47)$$

$$1 - \frac{1}{2\sqrt{\xi_k \gamma_k}} - \frac{1}{\gamma_k} \left[\exp\left(-\frac{1}{\xi_k} - \frac{\gamma_k}{4} + \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \sinh\left(\frac{\gamma_k}{4} - \sqrt{\frac{\gamma_k}{2\xi_k}}\right) \right]$$

Figure 2.8 illustrates gain curves for the $GEMMSE^1$ estimator for various values of ξ_k . It is interesting to note the dissimilarities in behavior between the $GEMMSE^1$ and $GEMMSE^1$ solutions. The increased attenuation applied by the former is due to the narrow peak of the exponential distribution, which shifts spectral estimates downward.

2.2.3.2 Assuming Exponential Noise Priors ($\zeta_n = 1, \nu_n = 1$)

In this section, a general MMSE solution is derived which assumes exponential noise spectral magnitude priors, with speech prior distributions constrained by $\zeta_x=1$, given as a function of $\nu_x \in \mathbb{N}_1$, where \mathbb{N}_1 is the set of natural numbers. In this case, Equation 2.39 can be simplified as

$$\hat{A}_k = \frac{\int_0^{R_k} A_k^{\nu_x} \exp(-(\beta_x - \beta_n) A_k) \partial A_k}{\int_0^{R_k} A_k^{\nu_x - 1} \exp(-(\beta_x - \beta_n) A_k) \partial A_k} \quad (2.48)$$

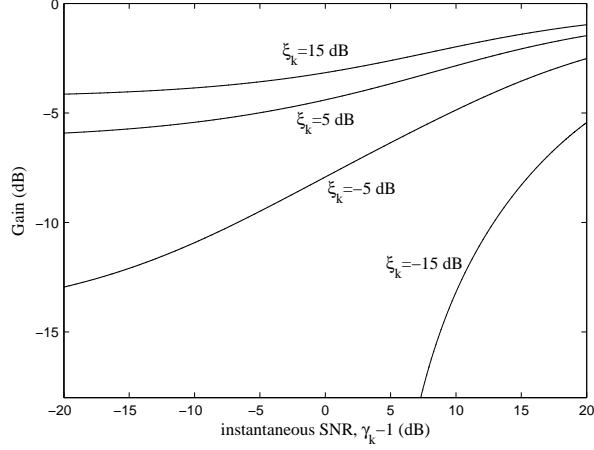


Figure 2.8: Gain curves for the \hat{G}_{GEMMSE}^1 STSA estimator for various values of ξ_k

The following identity, which is proven in Appendix A.4, is helpful in deriving the current spectral estimator

$$\int \tau^n \exp(c\tau) d\tau = \frac{\exp(c\tau)}{c} \sum_{k=0}^n \left[(-1)^k \frac{n!}{(n-k)!} \frac{\tau^{n-k}}{c^k} \right]. \quad (2.49)$$

Using Eqs. 2.48 and 2.49, the MMSE STSA estimator for exponential noise priors can be expressed as

$$G_{EMMSE}^{(\nu_x)}(\xi_k, \gamma_k) = \frac{1}{\mu_k} \left[\frac{(-1)^{\nu_x} \nu_x! + \exp(\mu_k) \sum_{m=0}^{\nu_x} \left((-1)^m \frac{\nu_x!}{(\nu_x-m)!} \mu^{\nu_x-m} \right)}{(-1)^{\nu_x-1} (\nu_x-1)! + \exp(\mu_k) \sum_{m=0}^{\nu_x-1} \left((-1)^m \frac{(\nu_x-1)!}{(\nu_x-m-1)!} \mu^{\nu_x-m-1} \right)} \right] \quad (2.50)$$

where

$$\mu_k = \sqrt{\frac{2\gamma_k}{\xi_k}} \left(\sqrt{\xi_k} - \sqrt{\frac{\nu_x(\nu_x-1)}{2}} \right). \quad (2.51)$$

Note that the gain function in Equation 2.50 is expressed as a function of ν_x , which allows flexibility in modeling the speech component. Note also that for $\mu_k = 0$, $G_{EMMSE}^{(\nu_x)}(\xi_k, \gamma_k)$ assumes an indeterminate form. L'Hopital's Rule can be used to evaluate $G_{EMMSE}^{(\nu_x)}(\xi_k, \gamma_k) |_{\mu_k=0}$, which is proven in Appendix A.5

$$G_{EMMSE}^{(\nu_x)}(\xi_k, \gamma_k) |_{\mu_k=0} = \frac{\nu_x}{\nu_x + 1}. \quad (2.52)$$

For the example case of $\nu_x=1$, the EMMSE solution reduces to

$$G_{EMMSE}^{(\nu_x=1)} = \frac{1}{\mu_k} \left(\frac{-1 + \exp(\mu_k)(-1 + \mu_k)}{1 + \exp(\mu_k)} \right), \quad (2.53)$$

Figure 2.9 provides gain curves for the $G_{EMMSE}^{(\nu_x)}$ estimator, for $\nu_x \in \{1, 2\}$. As can be observed, with increasing ν_x , the estimator generally provides decreased attenuation. This is due to the corresponding *a priori* speech distribution shifting its mode outward, thereby increasing the expected value of A_k .

2.2.4 MAP Estimation

This section presents a group of MAP spectral amplitude estimators assuming phase equivalence of speech and noise. Gaussian and Rayleigh noise priors are considered. As with MMSE estimation, higher order shape parameters offer valid solutions; however, higher order GGDs typically result in solutions which can not be expressed in closed form.

Applying the phase equivalence assumption to Eq. 1.27 yields

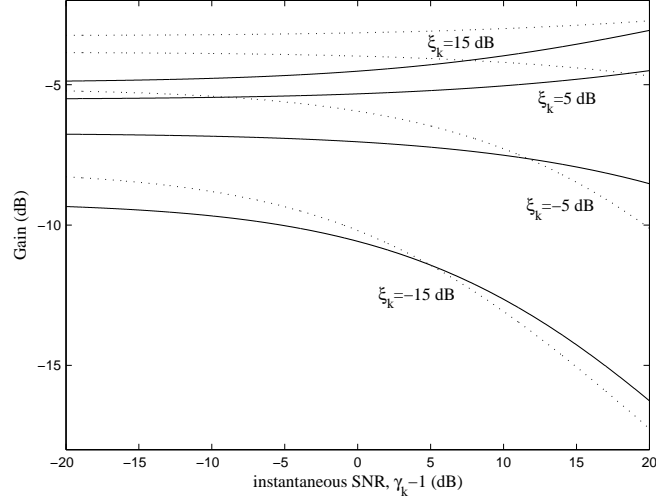


Figure 2.9: Gain curves for the $G_{EMMSE}^{(\nu_x)}$ STSA estimator for $\nu_x=1$ (solid line) and $\nu_x=2$ (dotted line)

$$\frac{\zeta_x \nu_x - 1}{A_k} - \frac{\zeta_n \nu_n - 1}{R_k - A_k} - \beta_x \zeta_x A_k^{\zeta_x - 1} + \beta_n \zeta_n (R_k - A_k)^{\zeta_n - 1} = 0 \quad (2.54)$$

Analyzing the 2^{nd} derivative of $C(A_k)$ with respect to A_k leads to

$$\begin{aligned} \frac{\partial^2}{\partial A_k^2} C(A_k) = & \quad (2.55) \\ & - \frac{\zeta_x \nu_x - 1}{A_k^2} - \frac{\zeta_n \nu_n - 1}{(R_k - A_k)^2} - \beta_x \zeta_x (\zeta_x - 1) A_k^{\zeta_x - 2} - \beta_n \zeta_n (\zeta_n - 1) (R_k - A_k)^{\zeta_n - 2} \end{aligned}$$

Equation 2.55 shows that $\frac{\partial^2}{\partial A_k^2} C(A_k)$ will be guaranteed negative, and the solution from Equation 2.54 valid, if each of the following inequalities holds

$$(i) \zeta_x \geq 1 \tag{2.56}$$

$$(ii) \zeta_n \geq 1$$

$$(iii) \zeta_x \nu_x \geq 1$$

$$(iv) \zeta_n \nu_n \geq 1$$

Equation 2.54 can assume negative values even if certain constraints in 2.56 are not met; however, these inequalities offer a simple check which encompasses the majority of valid speech and noise GGD shape parameters.

2.2.4.1 Assuming Gaussian Noise Priors ($\zeta_n=2, \nu_n=1/2$)

This section presents a family of MAP spectral magnitude estimators assuming Gaussian noise prior distributions. Separate solutions are derived for speech distributions constrained by ($\zeta_x=2, \nu_x \in \mathbb{R}$) and ($\zeta_x=1, \nu_x \in \mathbb{R}$).

The *G2MAP* estimator is derived from the assumption of Gaussian noise and speech distributions constrained by ($\beta_x=2, \nu_x \in \mathbb{R}$). In this case, Equation 2.54 reduces to

$$A_k^2 - \frac{\xi_k}{2\nu_x + \xi_k} R_k A_k - \frac{\xi_k (2\nu_x - 1)}{\gamma_k (2\nu_x + \xi_k)} = 0 \tag{2.57}$$

Applying the quadratic equation, and choosing the nonnegative root, yields

$$G_{G2MAP}^{(\nu_x)}(\xi_k, \gamma_k) = \frac{\xi_k + \sqrt{\xi_k^2 + 4(\xi_k/\gamma_k)(2\nu_x - 1)(2\nu_x + \xi_k)}}{2(2\nu_x + \xi_k)} \tag{2.58}$$

It is interesting to note that the $G_{G2MAP}^{(\nu_x)}$ estimator of Eq. 2.58 shows striking similarity to the MAP STSA estimator proposed by Wolfe and Godsil in [WG03]. Furthermore, the

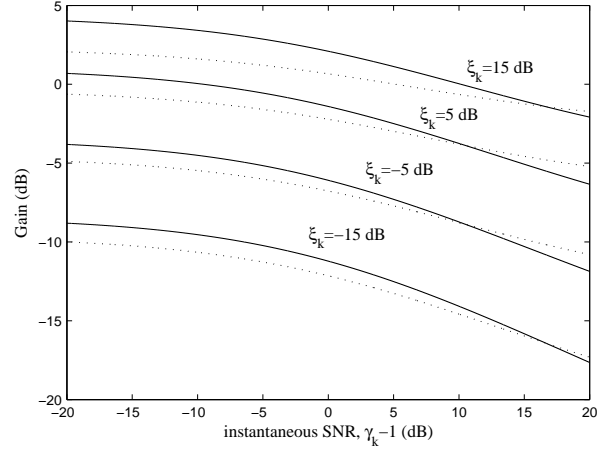


Figure 2.10: Gain curves for the $G_{G2MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=1$ (dotted line)

case of $\nu_x=1/2$ leads to the Wiener filter

$$G_{G2MAP}^{\nu_x=1/2}(\xi_k, \gamma_k) = \phi \quad (2.59)$$

Figure 2.10 illustrates gain curves for the $G_{G2MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=1$ (dotted line).

The steps involved in deriving MAP spectral estimators, which were previously followed to obtain the $G2MAP$ solution, are summarized as

1. Choose GGD size parameters corresponding to desired speech and noise prior distributions, and check for validity (Eq. 2.56).
2. Substitute GGD size parameters into the general solution for MAP estimation (Eq. 2.54).
3. Solve Eq. 2.54 for \hat{A}_k . Note that in certain cases, such as when $\frac{\partial}{\partial A_k} C(A_k)$ is linear or quadratic, this can be done in closed form. Other cases may rely on numerical approximations.
4. If multiple roots are obtained for \hat{A}_k choose the root which falls within the desired range $[0, \infty)$.
5. Substitute the definitions of the a priori and a posteriori SNRs (Eq. 1.7) into \hat{A}_k .

Following the previously outlined steps, the *G1MAP* solution is derived from the assumption of $(\beta_x = 1, \nu_x \in \mathbb{R})$

$$G_{G1MAP}^{(\nu_x)}(\xi_k, \gamma_k) = \frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}} + \sqrt{\left(\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}}\right)^2 + \frac{(\nu_x - 1)}{\gamma_k}} \quad (2.60)$$

Figure 2.11 illustrates gain curves for the $G_{G1MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=2.5$ (dotted line).

2.2.4.2 Assuming Rayleigh Noise Priors ($\zeta_n = 2, \nu_n = 1$)

This section presents a family of MAP spectral magnitude estimators assuming Rayleigh noise prior distributions. Equation 2.54 now reduces to

$$(\zeta_x \nu_x - 1) - \beta_x \zeta_x A_k^{\zeta_x} - \frac{A_k}{R_k - A_k} + 2\beta_n A_k (R_k - A_k) = 0. \quad (2.61)$$

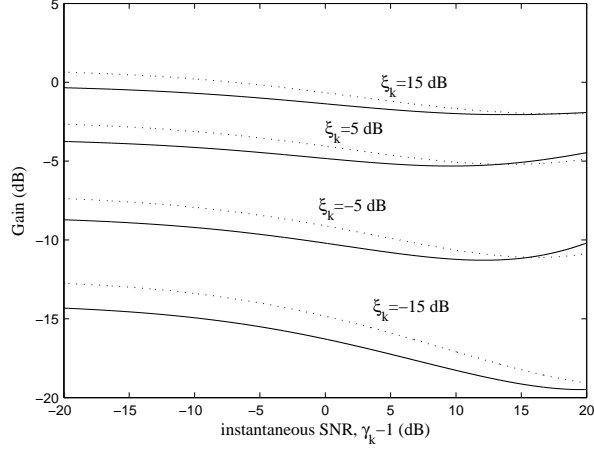


Figure 2.11: Gain curves for the $G_{G1MAP}^{(\nu_x)}$ STSA estimator for $\nu_x=2$ (solid line) and $\nu_x=2.5$ (dotted line)

It can be observed that the expression in Equation 2.61 is a $(\zeta_x + 1)^{th}$ -order polynomial. However, if the speech distribution shape parameters are constrained such that $\zeta_x \nu_x = 1$, the expression can be reduced to a ζ_x^{th} -order polynomial, and its roots can be obtained more efficiently. If Gaussian speech prior distributions are assumed, the *RGMAP* solution can be derived according to the steps outlined in Section 2.2.4.1

$$G_{RGMAP}(\xi_k, \gamma_k) = \frac{1}{2} \left(\frac{4\xi_k + 1}{2\xi_k + 1} \right) + \frac{1}{2} \sqrt{\left(\frac{4\xi_k + 1}{2\xi_k + 1} \right)^2 - \frac{4\xi_k}{\gamma_k} \left(\frac{2\gamma_k - 1}{2\xi_k + 1} \right)} \quad (2.62)$$

Instead if exponential speech priors are assumed, the *REMAP* estimator is derived as

$$G_{REMAP}(\xi_k, \gamma_k) = 1 - \frac{1}{2\sqrt{2\xi_k\gamma_k}} + \sqrt{\left(1 - \frac{1}{2\sqrt{2\xi_k\gamma_k}}\right)^2 + \left(\frac{1}{\sqrt{2\xi_k\gamma_k}} + \frac{1}{2\gamma_k} - 1\right)} \quad (2.63)$$

It is interesting to note the similarity between the proposed $G_{G1MAP}^{(\nu_x)}$ and $G_{REMAP}(\xi_k, \gamma_k)$

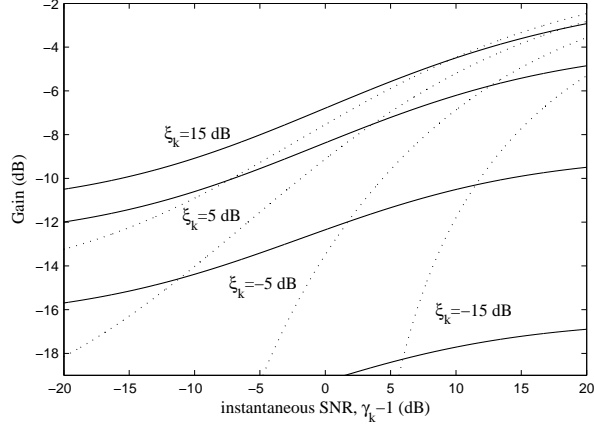


Figure 2.12: Gain curves for the G_{RGMAP} (solid line) and G_{REMAP} (dotted line) STSA estimators

solutions, and the MAP STSA estimator proposed by Lotter and Vary in [LV05]. Figure 2.12 illustrates gain curves for the G_{RGMAP} (solid line) and G_{REMAP} (dotted line) STSA estimators.

Table 2.3 provides a summary of the generalized STSA estimators derived in Sections 2.2.2-2.2.4.

2.2.5 Experimental Results

To assess the performance of the proposed speech enhancement methods, the Noizeus database is again used, and STSA estimators are again embedded into code from [Coh]. Table 2.4 provides quantitative results for a subset of the proposed STSA estimators, when applied to the Noizeus database [HL07]. Bold entries denote the best score for each metric at each noise condition. Results in Table 2.4 were obtained with a subtraction factor $\rho_k(\gamma_k, \theta_k)$ evaluated at $\theta_k = 7\pi/8$.

Table 2.4 shows the proposed STSA estimators to generally provide improved SSNRs, relative to the baseline MMSE method from [EM84], especially for inactive speech frames ($SSNR_N$). The $GGMMSE^1$ estimator, which provides the best scores, outper-

Table 2.3: STSA estimators derived in Sections 2.2.2, 2.2.3, and 2.2.4: Note that particular solutions are obtained by substituting into general solutions those statistical parameters corresponding to desired noise and speech priors.

Name	$G(\xi_k, \gamma_k)$
G_{ML}	$1 - \frac{1}{\sqrt{\gamma_k}} \left(\frac{\zeta_n \nu_n - 1}{\zeta_n \sqrt{\nu_n(\nu_n + 2 - \zeta_n)}} \right)^{1/\zeta_n}$
\hat{G}_{GGMMSE}^1	$\phi_k \left[1 - \frac{1}{\gamma_k} \exp \left(-\frac{\gamma_k(1 + \xi_k^2)}{4\xi_k(1 + \xi_k)} \right) \sinh \left(\frac{\gamma_k(\xi_k - 1)}{4\xi_k} \right) \right]$
\hat{G}_{GEMMSE}^1	$1 - \frac{1}{2\sqrt{\xi_k \gamma_k}} - \frac{1}{\gamma_k} \left[\exp \left(-\frac{1}{\xi_k} - \frac{\gamma_k}{4} + \sqrt{\frac{\gamma_k}{2\xi_k}} \right) \sinh \left(\frac{\gamma_k}{4} - \sqrt{\frac{\gamma_k}{2\xi_k}} \right) \right]$
$G_{EMMSE}^{(\nu_x)}$	$\frac{1}{\mu_k} \left(\frac{(-1)^{\nu_x} \nu_x! + \exp(\mu_k) \sum_{k=0}^{\nu_x} (-1)^k \frac{\nu_x!}{(\nu_x - k)!} \mu_k^{\nu_x - k}}{(-1)^{\nu_x - 1} (\nu_x - 1)! + \exp(\mu_k) \sum_{k=0}^{\nu_x - 1} (-1)^k \frac{(\nu_x - 1)!}{(\nu_x - k - 1)!} \mu_k^{\nu_x - k - 1}} \right)$ where $\mu_k = \sqrt{\frac{2\gamma_k}{\xi_k}} \left(\sqrt{\xi_k} - \sqrt{\frac{\nu_x(\nu_x - 1)}{2}} \right)$
$G_{G2MAP}^{(\nu_x)}$	$\frac{1}{2(2\nu_k + \xi_k)} \left(\xi_k + \sqrt{\xi_k^2 + 4(\xi_k/\gamma_k)(2\nu_x - 1)(2\nu_x + \xi_k)} \right)$
$G_{G1MAP}^{(\nu_x)}$	$\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}} + \sqrt{\left(\frac{1}{2} - \frac{\sqrt{\nu_x(\nu_x + 1)}}{2\sqrt{\gamma_k \xi_k}} \right)^2 + \frac{(\nu_x - 1)}{\gamma_k}}$
G_{RGMAP}	$\frac{1}{2} \left(\frac{4\xi_k + 1}{2\xi_k + 1} \right) + \frac{1}{2} \sqrt{\left(\frac{4\xi_k + 1}{2\xi_k + 1} \right)^2 - \frac{4\xi_k}{\gamma_k} \left(\frac{2\gamma_k - 1}{2\xi_k + 1} \right)}$
G_{REMAP}	$1 - \frac{1}{2\sqrt{2\xi_k \gamma_k}} + \sqrt{\left(1 - \frac{1}{2\sqrt{2\xi_k \gamma_k}} \right)^2 + \left(\frac{1}{\sqrt{2\xi_k \gamma_k}} + \frac{1}{2\gamma_k} - 1 \right)}$

Table 2.4: Segmental SNR scores for selected STSA estimators: Bold entries denote the best score for each metric at each noise condition.

Estimator	15 dB			10 dB			5 dB			0 dB		
	Δ	Δ_S	Δ_N	Δ	Δ_S	Δ_N	Δ	Δ_S	Δ_N	Δ	Δ_S	Δ_N
<i>MMSE</i> [EM84]	4.1	2.3	8.3	4.7	3.1	8.4	5.4	4.0	8.6	6.1	4.9	8.6
\hat{G}_{GGMMSE}^1	5.2	1.8	13.7	6.3	3.0	14.4	7.4	4.4	15.0	8.6	6.0	15.0
$G_{EMMSE}^{(\nu_x=1)}$	4.7	2.5	10.0	5.4	3.4	10.1	6.2	4.4	10.2	6.9	5.5	10.1
$G_{G2MAP}^{(\nu_x=1/2)}$	5.2	2.5	11.9	6.1	3.5	12.3	7.0	4.7	12.6	7.9	5.9	12.6
G_{RGMAP}	5.1	2.5	11.6	5.9	3.5	11.8	6.8	4.7	12.1	7.7	5.8	12.0
G_{REMAP}	4.9	2.5	10.7	5.6	3.4	10.9	6.4	4.5	11.0	7.2	5.6	10.9

forms that from [EM84] by approximately 1-2.5 dB in terms of SSNR across noise levels, and by approximately 5-6.5 dB in terms of $SSNR_N$.

Table 2.5 provides COSH measures for a subset of the proposed STSA estimators, when applied to the Noizeus database [HL07]. Bold entries denote the best score for each metric at each noise condition. It can be observed that the proposed estimators provide lower distortion measures than the MMSE solution from [EM84], across all noise levels. Furthermore, the best results are achieved by the \hat{G}_{GGMMSE}^1 estimator.

Informal listening tests show the proposed estimators included in Table 2.4 to provide a noticeable improvement in noise suppression, across noise types and levels, relative to the MMSE estimator from [EM84]. Speech enhancement at low noise levels (5 dB and 0 dB), did result in some apparent musical noise, especially for highly non-stationary, speech-shaped noise types, such as *babble* and *restaurant*. However, this musical noise was no more noticeable than that produced by [EM84].

Table 2.5: COSH Distortion Measures [JM76] for Selected STSA Estimators. Bold entries denote the best score for each metric at each noise condition. Results were obtained on the Noizeus database [HL07].

Estimator	(ζ_n, ν_n)	(ζ_x, ν_x)	15	10	5	0
$MMSE$ [EM84]	(2, 1)	(2, 1)	1.77	3.36	5.88	10.59
\hat{G}_{GGMMSE}^1	(2, 1/2)	(2, 1/2)	1.33	2.07	3.26	5.48
$G_{EMMSE}^{(\nu_x=1)}$	(1, 1)	(1, 1)	1.47	2.74	4.81	8.63
$G_{G2MAP}^{(\nu_x=1/2)}$	(2, 1/2)	(2, 1/2)	1.39	2.37	3.97	6.93
G_{RGMAP}	(2, 1)	(2, 1/2)	1.39	2.41	4.10	7.23
G_{REMAP}	(2, 1)	(1, 1)	1.45	2.58	4.48	7.95

Table 2.6: Word-accuracy rates for front-end short-time spectral amplitude estimation

SNR (dB)	20	15	10	5	0	Ave.
none	95.8	89.6	70.4	34.6	5.2	59.1
MMSE, $\nu=1$ (Eq. 2.1.2)	97.0	94.4	87.8	74.1	49.2	80.5
G2MAP (Eq. 2.58)	96.7	94.0	87.6	74.0	50.0	80.5
EMMSE (Eq. 2.53)	96.2	93.4	86.8	72.3	47.5	79.2

2.3 Speech Enhancement as a Method for Noise Robust ASR

Although speech enhancement is generally designed to improve the perceptual quality of speech signals degraded by acoustic noise, such techniques can also be expected to improve ASR since they provide the recognizer with an estimate of the underlying clean signal. Following the experimental procedure outlined in Section 1.3.3, a subset of STSA estimators proposed in this chapter are tested for the task of ASR. Table 2.6 provides word-accuracies for ASR with front-end speech enhancement. In Table 2.6, STSA estimation is shown to be a simple yet effective approach to front-end noise robust ASR.

This chapter proposes frameworks for deriving STSA estimators. Section 2.1 explores the use of GGD speech priors. Section 2.2 applies the assumption of phase equivalence of speech and noise. In each case, solutions are derived as functions of distribution shape

parameters. In many cases, it is shown that particular solutions reduce to well-known traditional estimators.

CHAPTER 3

Exploiting Temporal Correlation in Short-Time Spectral Amplitude Estimation

3.1 Statistical Framework

Traditional STSA estimation approaches determine time-frequency specific gains based on observed spectral amplitudes and a local estimate of noise statistics, to obtain an approximation of the spectral amplitude of the underlying clean speech ([MM80], [WG03], [EHH07]). Such a framework assumes successive windowed speech segments, and thus successive spectral components, to be statistically independent for the sake of mathematical simplicity [EM84]. Assumption of independent temporal components is generally not statistically valid, especially for overlapping analysis windows, and thereby fails to exploit correlation along the time axes.

This chapter introduces a model for the dynamic behavior of STSAs across time. The proposed model defines the statistical relationship between distinct spectral samples as a function of their relative time-frequency positions. The proposed model is used to derive the conditional distributions of observed spectral samples, based on sets of neighboring hidden clean STSAs.

3.1.1 Temporal Correlation of Clean STSAs

It is widely known that spectral speech data exhibits a high degree of correlation along the time axis [Ril89]. This characteristic has previously been exploited for various speech

processing tasks, such as spectral reconstruction for robust ASR ([BA10a], [KH10]). This section explores the inter-frame power correlation coefficients of speech spectral data through empirical analysis.

Assuming stationarity with respect to time, the STSA power correlation coefficient is defined as

$$\rho_k(\tau) \triangleq \frac{E[A_k(n)A_k(n+\tau)]}{\sigma_{X,k}^2} \quad (3.1)$$

Figure 3.1 provides an empirical study of power correlation coefficients, obtained from clean data in the Noizeus database [HL07]. The top panel illustrates first-order inter-frame power correlation coefficients, $\rho_k(1)$, as a function of frequency channel, k . It can be observed that successive power spectral components of clean speech exhibit a high degree of correlation, which generally lies above $\rho_k(\tau)=0.75$. It should be noted that small correlation coefficients for very high or very low channels is a result of attenuation applied at corresponding frequencies by the MIRS filter (see Section 1.2.6). The bottom panel illustrates inter-frame power correlation coefficients for example frequency channels, as a function of delay, τ . Visual inspection reveals a decaying trend, similar to an exponential relationship.

Figure 3.1 motivates the use of an exponential model for $\rho_k(\tau)$ with respect to time delay τ , leading to

$$\rho_k(\tau) = \lambda_k^{|\tau|}, \quad (3.2)$$

where λ_k denotes the rate parameter.

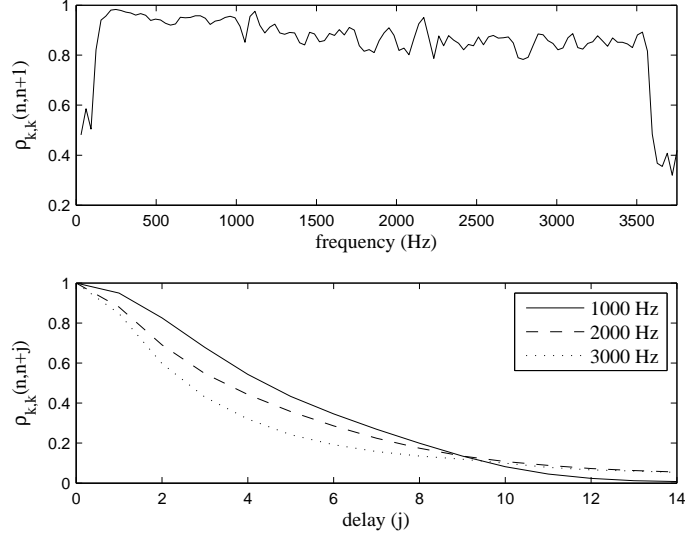


Figure 3.1: An empirical study of power correlation coefficients, obtained from the Noizeus database [HL07]. The top panel illustrates first-order inter-frame power correlation coefficients, $\varrho_k(1)$, as a function of channel index. The bottom panel illustrates inter-frame power correlation coefficients for example frequency channels, as a function of delay, τ .

3.1.2 Modeling the Dynamic Behavior of Clean STSAs with Respect to Time

This section presents a novel model for the dynamic behavior of short-time spectra of clean speech, which defines the statistical relationship between distinct samples as a function of their relative temporal positions. The model interprets underlying clean STSAs as hidden processes with inter-frame correlations defined by the distribution $p(A_k(n + \tau) | A_k(n))$. For the sake of mathematical simplicity, the following assumption is made regarding independence of conditional distributions of observed spectral components

$$p(A_k(n + \tau_1), A_k(n + \tau_2) | A_k(n)) = p(A_k(n + \tau_1) | A_k(n))p(A_k(n + \tau_2) | A_k(n)),$$

for $\tau_1 \neq \tau_2$ (3.3)

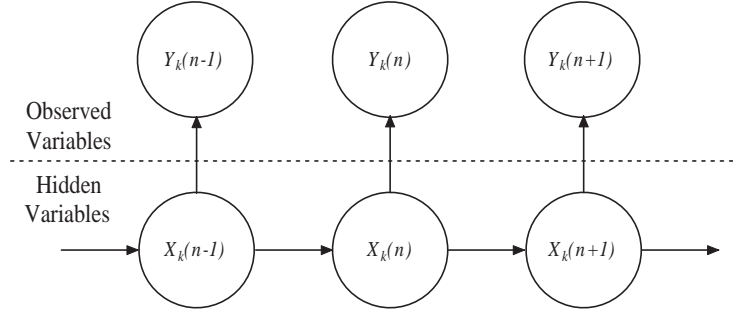


Figure 3.2: A graphical representation of the proposed model for the dynamic behavior of short-time spectral amplitudes of clean ($X_k(n)$) and observed ($Y_k(n)$) speech.

Observed STSAs are generated independently at each time-frequency component, as a function of the hidden clean STSA and the current noise statistics

$$\begin{aligned}
 & p(Y_k(n + \tau_1), Y_k(n + \tau_2) | A_k(n + \tau_1), A_k(n + \tau_2)) \\
 & = p(Y_k(n + \tau_1) | A_k(n + \tau_1)) p(Y_k(n + \tau_2) | A_k(n + \tau_2)), \text{ for } \tau_1 \neq \tau_2
 \end{aligned} \tag{3.4}$$

Figure 3.2 provides a graphical representation of the proposed model. Note that spectral observations are solely dependent on corresponding hidden variables.

For the proposed model the temporal dependency of clean speech STSAs is defined by the distribution $p(A_k(n + \tau) | A_k(n))$. A Bayesian approach results in

$$p(A_k(n + \tau) | A_k(n)) = \frac{p(A_k(n + \tau), A_k(n))}{p(A_k(n))}. \tag{3.5}$$

The joint probability of STSAs separated with respect to time is analogous to the estimation of correlated Rayleigh distributed fading channels in dual diversity communication systems [TJ00]. From [TJ00], the joint distribution of correlated Rayleigh random variables is given as

$$p(A_k(n+\tau), A_k(n)) = \frac{4A_k(n+\tau)A_k(n)}{\sigma_{X,k}^4(1-\varrho_k(\tau))} I_0 \left(\frac{2\sqrt{\varrho_k(\tau)}A_k(n+\tau)A_k(n)}{\sigma_{X,k}^2(1-\varrho_k(\tau))} \right) \quad (3.6)$$

$$\times \exp \left(-\frac{A_k^2(n+\tau) + A_k^2(n)}{\sigma_{X,k}^2(1-\varrho_k(\tau))} \right),$$

Substitution of Eq. 3.6 along with the Rayleigh marginal distribution of A_k , into Eq. 3.5, results in the conditional distribution

$$p(A_k(n+\tau) | A_k(n)) = \frac{2A_k(n+\tau)}{\sigma_{X,k}^2(1-\varrho_k(\tau))} I_0 \left(\frac{2\sqrt{\varrho_k(\tau)}A_k(n+\tau)A_k(n)}{\sigma_{X,k}^2(1-\varrho_k(\tau))} \right) \quad (3.7)$$

$$\times \exp \left(-\frac{A_k^2(n+\tau) + \varrho_k(\tau)A_k^2(n)}{\sigma_{X,k}^2(1-\varrho_k(\tau))} \right),$$

which defines the time dependency of clean STSAs. It is interesting to note that the distribution defined by Eq. 3.7 follows a Rician distribution. Additionally, for uncorrelated data, i.e. $\varrho_k(\tau) = \delta_\tau$ where δ_τ is the Kronecker delta function, Eq. 3.7 reduces to the univariate Rayleigh distribution of Eq. A.1.

3.2 Conditional Probabilities Based on Multiple Observed Spectral Components

Traditional STSA estimation approaches infer underlying clean speech amplitudes based on single corresponding observed spectral components. That is, such STSA techniques generally base optimal solutions on the statistical distribution $p(Y_k(n) | A_k(n))$. In this section, conditional probabilities are derived based on multiple spectral observations. Letting the set of neighboring observed spectral components be denoted by

$$\mathbf{Y}_k(n) = \{Y_k(n - t_1), \dots, Y_k(n + t_2)\}, \text{ for } t_1, t_2 \geq 0 \quad (3.8)$$

the conditional distribution $p(\mathbf{Y}_k(n) | A_k(n))$ is derived. Similar to Eq. 3.8, γ_k is defined as the corresponding set of neighboring observed *a posteriori* SNRs. Finally, let \mathcal{T} represent set of time indices included in $\mathbf{Y}_k(n)$

$$\mathcal{T} = \{-t_1, \dots, t_2\} \quad (3.9)$$

Utilizing the model of dynamic STSA statistical behavior from section 3.1.2, and the assumption from Eq. 3.3, the following distribution can be obtained

$$P(\mathbf{Y}_k(n) | A_k(n)) = \prod_{\tau \in \mathcal{T}} p(Y_k(n + \tau) | A_k(n)) \quad (3.10)$$

where the conditional probability $p(Y_k(n + \tau) | A_k(n))$ is determined via marginalization

$$\begin{aligned} p(Y_k(n + \tau) | A_k(n)) & \quad (3.11) \\ &= \int_0^\infty p(Y_k(n + \tau) | A_k(n + \tau)) p(A_k(n + \tau) | A_k(n)) \partial A_k(n + \tau). \end{aligned}$$

Substitution of Eqs. 1.14 and 3.5, and the large-value approximation of $I_0(\cdot)$ from Eq. A.20, results in

$$\begin{aligned}
& p(Y_k(n + \tau) | A_k(n)) \tag{3.12} \\
&= (1 + \operatorname{erfc}(\mu_k(\tau))) \frac{\exp\left(-\frac{1}{\hat{\sigma}_k^2(\tau)} \left(\sqrt{\varrho_k(\tau)} A_k(n) - R_k(n + \tau)\right)^2\right)}{2\pi \sqrt{\pi \hat{\sigma}_k^2(\tau) \rho_k^{1/2}(\tau) A_k(n) R_k(n + \tau)}}
\end{aligned}$$

where erfc denotes the complementary Gauss error function (see Eq. A.13) and where

$$\hat{\sigma}_k^2(\tau) \triangleq \sigma_{N,k}^2 + \sigma_{X,k}^2 (1 - \varrho_k(\tau)) \tag{3.13}$$

$$\mu_k(\tau) \triangleq \frac{\xi_k (1 - \varrho_k(\tau)) R_k(n + \tau) + \sqrt{\varrho_k(\tau)} A_k}{\sqrt{\xi_k (1 - \varrho_k(\tau)) \hat{\sigma}_k^2(\tau)}} \tag{3.14}$$

Extensive numerical experimentation showed the term $\operatorname{erfc}(\mu_k(\tau))$ to have converged to unity for the strong majority of cases, leading to the approximation $\operatorname{erfc}(\mu_k(\tau)) \approx 1$. Applying Eq. A.20 to Eq. 3.12 results in

$$\begin{aligned}
& p(Y_k(n + \tau) | A_k(n)) \tag{3.15} \\
&= \frac{1}{\pi \hat{\sigma}_k^2(\tau)} I_0 \left(\frac{2\sqrt{\varrho_k(\tau)} A_k(n) R_k(n + \tau)}{\hat{\sigma}_k^2(\tau)} \right) \exp \left(-\frac{\varrho_k(\tau) A_k^2(n) + R_k^2(n + \tau)}{\hat{\sigma}_k^2(\tau)} \right)
\end{aligned}$$

Note that for fully correlated STSAs, i.e. $(\varrho_k(\tau) = 1, \forall \tau)$, Eq. 3.15 reduces to the conditional distribution of Eq. 1.14. Conversely, for uncorrelated STSAs, i.e. $\rho_k(\tau) = \delta_\tau$, Eq. 3.15 assumes the marginal distribution of $Y_k(n)$ from Eq. 1.4. It is worth noting that Eq. 3.15 corresponds to the following conditional distribution of observed spectral amplitudes

$$\begin{aligned}
& p(R_k(n + \tau) | A_k(n)) \tag{3.16} \\
& = \frac{2R_k(n + \tau)}{\hat{\sigma}_k^2(\tau)} I_0 \left(\frac{2\sqrt{\varrho_k(\tau)} A_k(n) R_k(n + \tau)}{\hat{\sigma}_k^2(\tau)} \right) \exp \left(-\frac{\varrho_k(\tau) A_k^2(n) + R_k^2(n + \tau)}{\hat{\sigma}_k^2(\tau)} \right)
\end{aligned}$$

which is Rice-distributed. The Rician distribution was originally derived in the context of describing the envelope of a sinusoid in the presence of bandlimited additive noise [MW71]. The distribution is thereby expressed conveniently as a function of the ratio of the square of the signal amplitude to the noncentral second moment of the noise, i.e. the SNR. Analogously, the distribution of Eq. 3.16 can be interpreted as the detection of an unknown amplitude in the presence of uncertain information. However, in the latter case, the "SNR" is given by

$$\text{"SNR"} = \frac{\varrho_k(\tau) A_k^2(n)}{\sigma_{N,k}^2 (1 + \xi_k (1 - \varrho_k(\tau)))} \tag{3.17}$$

and is the result of both additive noise and correlation coefficients less than unity.

3.3 Maximum a Posteriori STSA Estimation: the CB-MAP Estimator

In this section, an optimal STSA estimator is derived based on groups of neighboring observed spectral components, utilizing statistical distributions derived in Section 3.2. A novel MAP solution is presented, and shown to be a generalized version of that proposed by Wolfe and Godsil [WG03].

Using the problem formulation from Section 1.2.4, the MAP solution incorporating neighboring observed spectral components during the estimation process is given as

$$\hat{A}_k(n) = A_k(n) \text{ such that} \quad (3.18)$$

$$(i) \frac{\partial}{\partial A_k(n)} \log p(\mathbf{Y}_k(n) | A_k(n)) p(A_k(n)) = 0$$

$$(ii) \frac{\partial^2}{\partial A_k^2(n)} \log p(\mathbf{Y}_k(n) | A_k(n)) p(A_k(n)) < 0$$

Applying Eq. A.20 to Eq. 3.18, constraint (i) leads to

$$\sum_{\tau \in \mathcal{T}} \left[\frac{\varrho_k(\tau)}{\hat{\sigma}_k^2(\tau) / \sigma_{N,k}^2} A_k^2(n) - \frac{\sqrt{\varrho_k(\tau)} R_k(n+\tau)}{\hat{\sigma}_k^2(\tau) / \sigma_{N,k}^2} A_k(n) + \frac{\sigma_{N,k}^2}{4} \right] - \frac{\sigma_{N,k}^2}{2} + \frac{A_k^2(n)}{\xi_k} = 0 \quad (3.19)$$

Similarly, Eq. 3.18 constraint (ii) leads to

$$\sum_{\tau \in \mathcal{T}} \left[-\frac{1}{2A_k^2(n)} - \frac{2\varrho_k(\tau)}{\hat{\sigma}_k^2(\tau)} \right] - \frac{1}{A_k^2(n)} - \frac{2}{\sigma_{N,k}^2} < 0 \quad (3.20)$$

which guarantees the MAP solution to be valid for $A_k(n) > 0$.

Applying the quadratic equation to Eq. 3.19 results in the proposed correlation-based maximum a posterior (CB-MAP) STSA gain function

$$G_{CB-MAP}(\xi_k(n), \gamma_k(n)) = \frac{\xi_k(n) H_\tau(\xi_k(n), \gamma_k(n))}{2 \left(1 + \xi_k(n) \hat{H}_\tau(\xi_k(n), \gamma_k(n)) \right)} \quad (3.21)$$

$$\times \left[1 + \sqrt{1 - \frac{(|\mathcal{T}| - 2) \left(1 + \xi_k(n) \hat{H}_\tau(\xi_k(n), \gamma_k(n)) \right)}{\xi_k(n) \gamma_k(n) H_\tau^2(\xi_k(n), \gamma_k(n))}} \right]$$

where $|\mathcal{T}|$ denotes the cardinality of \mathcal{T} , and the term

$$H_{\tau}(\xi_k(n), \gamma_k(n)) = \sum_{\tau \in \mathcal{T}} \frac{\sqrt{\varrho_k(\tau)}}{1 + \xi_k(n)(1 - \varrho_k(\tau))} \sqrt{\frac{\gamma_k(n + \tau)}{\gamma_k(n)}} \quad (3.22)$$

serves as an embedded low-pass FIR filter in the time-frequency domain. Specifically, $H_{\tau}(\xi_k(n), \gamma_k(n))$ applies tap values in time-frequency space proportional to the square root of the ratio of *a posteriori* SNRs. It should be noted that the presence of $H_{\tau}(\xi_k(n), \gamma_k(n))$ is similar in notion to optimal smoothing presented in [Mar01]. The term

$$\hat{H}_{\tau}(\xi_k(n), \gamma_k(n)) = \sum_{\tau \in \mathcal{T}} \frac{\varrho_k(\tau)}{1 + \xi_k(n)(1 - \varrho_k(\tau))} \quad (3.23)$$

serves as a normalization factor which accounts for the value of λ_k and for the cardinality of \mathcal{T} .

It is interesting to observe the behavior of $H_{\tau}(\xi_k(n), \gamma_k(n))$ for certain special cases. If the underlying STSA data is completely uncorrelated, i.e. $\varrho_k(\tau) = \delta_{\tau}$, $H_{\tau}(\xi_k(n), \gamma_k(n))$ is equivalent to

$$H_{\tau}(\xi_k(n), \gamma_k(n)) |_{\varrho_k(\tau) = \delta_{\tau}} = \delta_{\tau} \quad (3.24)$$

This follows intuitively, since for uncorrelated data, neighboring observations do not introduce useful information. In the case where the *a priori* SNR is high, $H_{\tau}(\xi_k(n), \gamma_k(n))$ is again equivalent to

$$H_{\tau}(\xi_k(n), \gamma_k(n)) |_{\xi_k \rightarrow \infty} = \delta_{\tau} \quad (3.25)$$

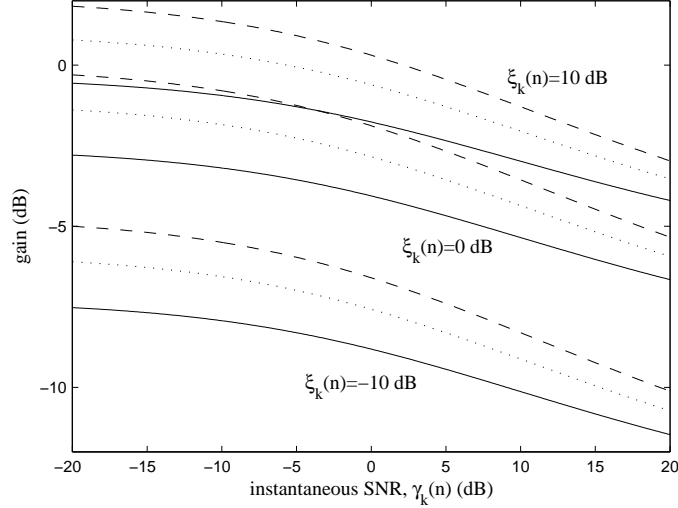


Figure 3.3: Gain curves for the proposed correlation-based maximum a posteriori STSA estimator for $\mathcal{T} = \{-1, 0\}$ and for $\gamma_k(n-1) = -5$ dB (solid line), $\gamma_k(n-1) = 5$ dB (dotted line), and $\gamma_k(n-1) = 10$ dB (dashed line). For illustrative purposes, $\lambda_k = 0.95$.

This follows intuitively as well, since for high values of the *a priori* SNR, observed spectral components can be considered dependable, and it becomes unnecessary to consider neighboring data. It should be noted that the proposed MAP solution is a generalized version of that presented by Wolfe and Godsil in [WG03], and the two are equivalent for the trivial case when $\mathcal{T} = \{0\}$.

Figure 3.3 provides gain curves for the CB-MAP estimator with $\mathcal{T} = \{-1, 0\}$. As an illustrative example, the temporal power correlation coefficient rate parameter is set as $\lambda_k = 0.95$, and the *a priori* SNR is constrained as $\xi_k(n) = \gamma_k(n) - 1$. Figure 3.3 displays the effect of neighboring observed spectral components on estimation of the current STSA. For example, if $\gamma_k(n-1)$ is observed to be small, the CB-MAP gain function will apply a high level of attenuation. Conversely, if $\gamma_k(n-1)$ is observed to be large, the CB-MAP estimator will decrease the level of attenuation.

3.4 Estimation Under Speech Presence Uncertainty: the CB-SPP filter

With motivation similar to that of Section 3.3, this section derives the correlation-based speech presence probabilities (CB-SPP) filter using GLRs which account for neighboring observed spectral components

$$\tilde{\Lambda}(\xi_k(n), \gamma_k(n)) \triangleq \eta_k \frac{p(\mathbf{Y}_k(n) | q_k(n) = H_1)}{p(\mathbf{Y}_k(n) | q_k(n) = H_0)}, \quad (3.26)$$

Using the statistical independence assumption of Eq. 3.3, $\tilde{\Lambda}(\xi_k(n), \gamma_k(n))$ can be expressed as

$$\tilde{\Lambda}(\xi_k(n), \gamma_k(n)) = \eta_k \frac{\prod_{\tau \in \mathcal{T}} p(Y_k(n + \tau) | q_k(n) = H_1)}{\prod_{\tau \in \mathcal{T}} p(Y_k(n + \tau) | q_k(n) = H_0)}. \quad (3.27)$$

It was observed in studies such as [Coh05], [GBM08], and [BA10b] that active and inactive time-frequency speech spectral components generally occur in salient segments. In these studies, the low probability of transition between states H_0 and H_1 is exploited to increase the accuracy of speech presence probability masks. Here, similar motivation is used to derive an improved GLR. Eq. 3.27 is comprised of the product of conditional probabilities, each of which can be decomposed using Bayes rule

$$p(Y_k(n + \tau) | q_k(n) = H_s) = \sum_{t=0}^1 [P(q_k(n + \tau) = H_t | q_k(n) = H_s) p(Y_k(n + \tau) | q_k(n + \tau) = H_t)], \text{ for } s \in \{0, 1\} \quad (3.28)$$

The term $p(Y_k(n + \tau) | q_k(n + \tau) = H_s)$ is determined from the marginal distributions of Eq. 1.4. The term $P(q_k(n + \tau) = H_t | q_k(n) = H_s)$ can be evaluated via a combi-

natorial approach by enumerating all possible binary paths between $q_k(n) = H_s$ and $q_k(n + \tau) = H_s$. Let a_{11} and a_{00} denote the self-transition probabilities for states H_1 and H_0 , respectively, along the temporal axis

$$a_{11} = P(q_k(n) = H_1 | q_k(n-1) = H_1) \quad (3.29)$$

$$a_{00} = P(q_k(n) = H_0 | q_k(n-1) = H_0) \quad (3.30)$$

The term p_τ^s is introduced as the probability of self-transition in state H_s , for time-frequency components with temporal separation τ

$$\begin{aligned} p_\tau^s &\triangleq P(q_k(n + \tau) = H_s | q_k(n) = H_s) \\ &= \left[\sum_{h=0}^{\lfloor |\tau|/2 \rfloor} \binom{|\tau|}{2h} (a_{ss})^{|\tau|-2h} (1 - a_{11})^h (1 - a_{00})^h \right] \end{aligned} \quad (3.31)$$

Note that transitional statistic p_τ^s is independent of the current signal, and can therefore be determined *a priori*. Figure 3.4 provides transitional probabilities p_τ^1 for example cases of a_{11} . For illustrative purposes, probabilities are determined with the constraint $a_{11}=a_{00}$.

Using Eq. 3.31, the improved GLR of Eq. 3.27 can be expressed as

$$\tilde{\Lambda}(\xi_k(n), \gamma_k(n)) = \eta_k \frac{\prod_{\tau \in \mathcal{T}} [1 - p_\tau^1 + p_\tau^1 \Lambda(\xi_k(n), \gamma_k(n))]}{\prod_{\tau \in \mathcal{T}} [p_\tau^0 + (1 - p_\tau^0) \Lambda(\xi_k(n), \gamma_k(n))]} \quad (3.32)$$

For the trivial case of $\mathcal{T}=\{0\}$, the proposed GLR reduces to that presented in [EM84].

Figure 3.5 provides gain curves for the CB-SPP filter. It can be observed that the level of attenuation applied by the CB-SPP filter increases as $\gamma_k(n-1)$ decreases.

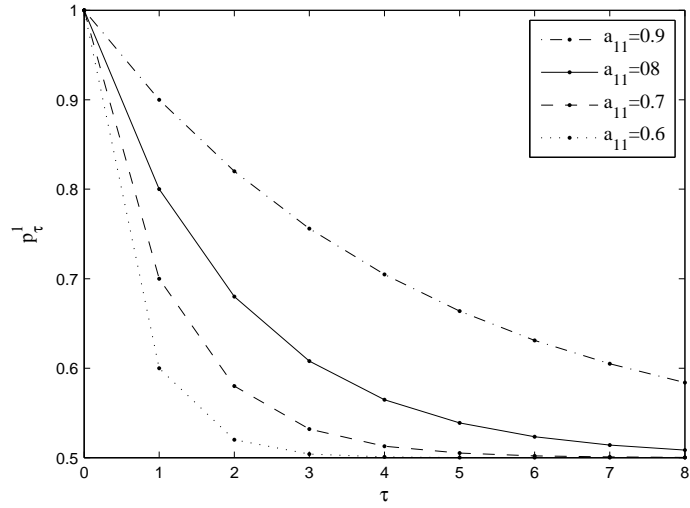


Figure 3.4: Transitional statistics for speech presence probability masks, as defined by Eq. 3.31: As an illustrative example, probabilities are determined with the constraint $a_{11}=a_{00}$.

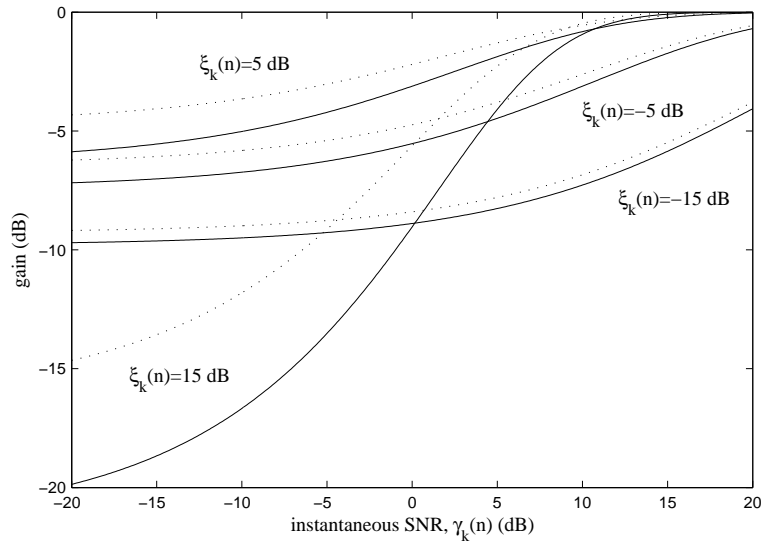


Figure 3.5: Gain curves for the correlation-based speech presence probability filter for $\mathcal{T}=\{-1, 0\}$, and for $\gamma_k(n-1)=10$ dB (dotted line) and $\gamma_k(n-1)=8$ dB (solid line). For illustrative purposes, $\lambda_k=0.95$ and $\eta_k=0.33$.

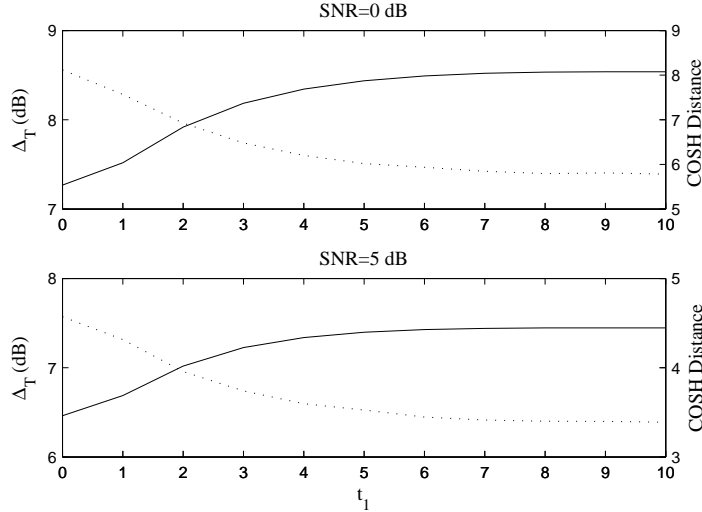


Figure 3.6: Improvements in SSNR_T (solid line) and COSH distance (dotted line) for the proposed CB-MAP estimator, as a function of t_1 . Note that for $t_1=0$, the CB-MAP estimator reduces to that proposed in [WG03].

3.5 Experimental Results

The Noizeus database is used to assess the performance of the speech enhancement methods. Figure 3.6 illustrates the effect of t_1 on resulting quality, in terms of differential SSNR_T and COSH distance. Speech quality for the CB-MAP estimator is observed to improve with the use of neighboring spectral observations. It can be observed that each measure improves as t_1 increases, and each seemed to converge at approximately $t_1 \approx 5$. Furthermore, the improvement due to exploiting temporal correlation is more pronounced for less favorable acoustic conditions. This is due to $H_\tau(\xi_k(n), \gamma_k(n))$ approaching δ_τ as the *a priori* SNR increases, and thereby having less effect on the CB-MAP estimator.

Table 3.2 provides speech enhancement results for the proposed STSA estimator. It can be observed that the integration of neighboring observed spectral components during STSA estimation improves SSNR scores. Specifically, proposed methods provide increased attenuation during inactive speech frames while maintaining similar speech quality for active speech frames. It is also shown that the use of correlation-based SPP filter

Table 3.1: Numerical parameters for STSA estimation methods described in Sections 3.3 and 3.4

Parameter	Value	Description
λ_k	0.15	Power correlation coefficient with respect to time
η	0.33	Ratio of a priori SPP probabilities
a_{11}	0.75	Transitional probability of states with respect to time
a_{00}	0.75	Transitional probability of states with respect to time

Table 3.2: Segmental SNR scores for proposed STSA estimators

t_1	15 dB			10 dB			5 dB			0 dB		
	Δ_T	Δ_S	Δ_N	Δ_T	Δ_S	Δ_N	Δ_T	Δ_S	Δ_N	Δ_T	Δ_S	Δ_N
without CB-SPP filter												
0	4.9	2.5	10.6	5.6	3.4	10.8	6.5	4.5	11.1	7.3	5.6	11.0
2	5.2	2.5	11.9	6.1	3.5	12.3	7.0	4.7	12.7	7.9	5.9	12.6
5	5.3	2.2	13.0	6.3	3.4	13.6	7.4	4.7	14.1	8.5	6.1	14.0
with CB-SPP filter												
0	5.3	2.4	12.6	6.2	3.4	13.0	7.3	4.7	13.5	8.2	6.1	13.4
2	5.3	2.3	12.9	6.3	3.4	13.5	7.3	4.7	13.9	8.3	6.1	13.9
5	5.3	2.1	13.2	6.3	3.3	13.8	7.4	4.6	14.5	8.5	6.1	14.3

generally improves SSNR-based scores.

Table 3.3 provides speech enhancement results in terms of the COSH distance [JM76]. The CB-MAP estimator is again shown to improve the quality of enhanced speech for $t_1 > 0$. The use of the CB-SPP filter is shown to further improve COSH distance in most cases.

Informal listening tests prove consistent with quantitative results of Table 3.2. The proposed solution provides an apparent increase in noise suppression as t_1 is increased. The use the SPP filter further increases the attenuation of noise. In many cases, the CB-MAP estimator reduces the level of perceptual musical artifacts. This may be due to its ability to incorporate sets of neighboring spectral components during estimation, thus

Table 3.3: COSH distance [JM76] for proposed STSA estimators

t_1	15 dB	10 dB	5 dB	0 dB
without CB-SPP filter				
0	1.47	2.69	4.57	8.12
2	1.39	2.37	3.96	6.92
5	1.33	2.16	3.53	6.02
with CB-SPP filter				
0	1.33	2.25	3.69	6.40
2	1.35	2.22	3.64	6.28
5	1.32	2.13	3.42	5.85

smoothing the effect of rapidly appearing residual tones.

This chapter presents short-time spectral amplitude (STSA) estimation exploiting temporal correlation of spectral speech data. A novel statistical model for the dynamic behavior of clean speech is derived. Using the proposed model of dynamic behavior of clean STSAs, a MAP solution to STSA estimation is proposed, which accounts for neighboring spectral observations. Additionally, a novel CB-SPP filter is derived which utilizes the model of speech dynamics.

CHAPTER 4

Improved Speech Presence Probabilities Using HMM-Based Inference

This chapter proposes a framework for determining improved speech presence probabilities using statistical inference. Motivation is provided for the assumption of standard SPPs as observations of channel-specific 2-state Markov models, wherein the states represent active and inactive spectral speech regions, respectively. Corresponding steady-state transitional statistics are set to capture the well-known temporal correlation of spectral speech data, and observation statistics can be modeled based on the effect of additive acoustic noise on resulting SPP values. Once underlying models have been parameterized, improved speech presence probabilities can be estimated by applying traditional HMM-based decoding techniques to standard SPPs. The decoding method in this section is robust to standard SPP filters obtained from various statistical models. However, in this study results are provided for the method proposed in [EM84] (Eq. 1.34), arising from Gaussian models.

Traditional SPPs (as described in Section 1.2.5) are based solely on the current frame, and do not exploit the well-known temporal correlation present in spectral speech data. This section presents a novel algorithm for determining improved SPPs. The algorithm can operate in two modes: utilizing past and present observations, or utilizing past, present, and future observations, to fully take advantage of inter-frame correlation.

4.1 Interpreting SPPs as Observations of Channel-Specific 2-State Models

SPP masks derived from speech in favorable acoustic environments reveal reliable speech components which tend to occur in salient segments, which is due in part to the well-known temporal correlation of time-frequency speech data. Additionally, in favorable acoustic environments, SPPs tend towards binary masks, i.e. SPPs assume either $P(H_1)$ or 1. This can be shown by examining Equation 1.32 for extreme values of the *a priori* SNR. When the speech component is zero, and the additive noise is very small, $\xi_k(n)=0$. Furthermore:

$$\Lambda_k(n)|_{\xi_k(n)=0} = \frac{P(H_1)}{P(H_0)} \Rightarrow P(H_1|Y_k(n))|_{\xi_k(n)=0} = P(H_1). \quad (4.1)$$

Conversely, when the magnitude of the spectral speech component is much larger than that of the additive noise, then $\xi_k(n) \rightarrow \infty$. In this case

$$\Lambda_k(n)|_{\xi_k(n) \rightarrow \infty} = \infty \Rightarrow P(H_1|Y_k(n))|_{\xi_k(n) \rightarrow \infty} = 1. \quad (4.2)$$

Thus, in the presence of very low background noise, SPP masks can be assumed binary. Furthermore, it is this binary mask that is interpreted to contain "true" speech presence probabilities. However, the low noise case represents oracle information, and is not accessible in realistic speech processing systems; instead speech signals tend to be corrupted by higher levels of background noise. Therefore, in determining improved SPPs, the underlying state is inferred, i.e. $P(H_1)$ or 1, for each spectro-temporal location given "noisy" standard SPP observations. Note that in order to simplify notation, posterior

probabilities obtained from traditional SPP methods are referred to as $\tau_k(n)$

$$\tau_k(n) = P(H_1|Y_k(n)). \quad (4.3)$$

By interpreting standard SPP masks as observations of channel-specific 2-state models, true binary masks can be estimated via traditional inference techniques. One such family of methods involves HMM-based decoding of noisy information [Rab89].

4.2 HMM-Based Mask Decoding

In order to apply HMM-based inference techniques, these statistical parameters must be determined for the hidden two-state model, H_i . Transitional statistics are set to capture the temporal correlation present in spectral speech data. In this study, a_{ij} will denote the probability of transition between H_i and H_j , for $i, j \in \{0, 1\}$, $1 \leq m \leq N_k$. The steady-state probability of state H_i can be obtained from transitional statistics as

$$P(H_i) = \frac{\sum_{j=0, j \neq i}^1 a_{ji}}{\sum_{g=0}^1 \sum_{h=0, h \neq g}^1 a_{gh}}. \quad (4.4)$$

Characterizing observation statistics for speech activity models H_i involves studying the effect of acoustic noise on corresponding speech presence probabilities. The distribution of observed reliability measures conditioned on underlying speech activity states is defined as

$$b_k^i(\tau_k(n)) = p(\tau_k(n) | H_i), \text{ for } 0 \leq \tau_k(n) \leq 1. \quad (4.5)$$

The relationship between additive acoustic noise in the spectral domain and the resulting inaccuracy of SPPs is difficult to express in closed form. Instead, statistical tools are required to model the distribution of noisy SPPs about their underlying binary values, as

a function of estimated additive acoustic noise.

It follows intuitively that the distribution of SPPs should reveal a global peak at the true binary value, and should display a monotonically decreasing probability density function (pdf) directly related to the distance from the underlying value. Considering these constraints, and due to their mathematical efficiency, state-conditional observation distributions are modeled as raised cosine distributions

$$\begin{aligned} b_k^0(\tau_k(n)) &= \left(\frac{1}{1 - P(H_1)} \right) \left(1 + \phi_k^0 \cos \left(\frac{\pi(\tau_k(n) - P(H_1))}{1 - P(H_1)} \right) \right), \\ b_k^1(\tau_k(n)) &= \left(\frac{1}{1 - P(H_1)} \right) \left(1 + \phi_k^1 \cos \left(\frac{\pi(1 - \tau_k(n))}{1 - P(H_1)} \right) \right), \end{aligned} \quad (4.6)$$

for $p(H_1) \leq \tau_k(n) \leq 1$.

As can be observed from the expressions in Equation 4.6, the parameter $\phi_k^i \in [0, 1]$ controls the effect of the sinusoidal component on the overall observation statistics. It follows intuitively that ϕ_k^i should be set to capture the estimated accuracy of observed channel-specific SPPs. That is, for clean spectro-temporal components, SPPs are determined with a high degree of accuracy, and thus observed close to their corresponding underlying binary states, i.e. $P(H_1)$ or 1. In this case, ϕ_k^i should be set close to 1. Conversely, for noisy spectro-temporal components, a higher degree of confusability is introduced into the estimation of SPPs, and observed SPPs may vary from their underlying binary states. In this case, ϕ_k^i should be set close to 0.

Considering the relationship between additive noise and the statistical parameter ϕ_k^i , it is proposed to be a function of the *a priori* SNR, ξ_k , of the given spectro-temporal component. Specifically, ϕ_k^i is proposed to be determined as

$$\phi_k^i(n) = \left(\frac{\xi_k(n)}{\xi_k(n) + \kappa^i} \right)^2. \quad (4.7)$$

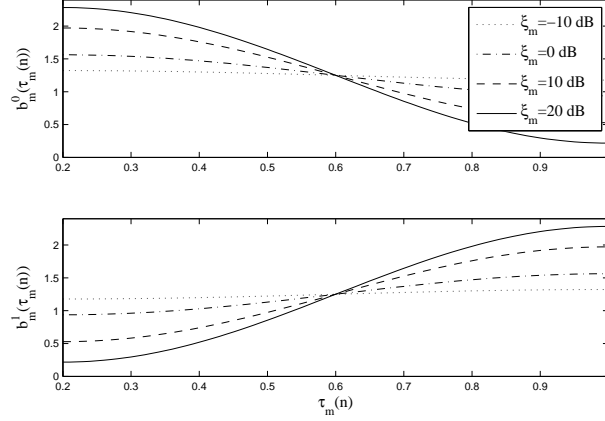


Figure 4.1: Example probability distribution functions $b_k^0(\tau_k(n))$ and $b_k^1(\tau_k(n))$ for various values of $\xi_k(n)$. For this example, $P(H_1)=0.2$, and $\kappa^i=1.0$.

Here, κ^i is an empirically tuned parameter. Figure 4.1 provides example probability distribution functions $b_k^0(\tau_k(n))$ and $b_k^1(\tau_k(n))$ for various values of $\xi_k(n)$. For this example, $p(H_1)=0.2$. As can be interpreted, when the *a priori* SNR is high (e.g. $\xi_k=20$ dB), the corresponding distribution decays rapidly. However, when the *a priori* SNR is low (e.g. $\xi_k=-10$ dB), the corresponding distribution decreases slowly, similar to a uniform pdf.

This chapter proposes improved speech presence probabilities, $\hat{\tau}_k(n)$, based on the set of traditional SPPs, $\{\tau_k(1), \dots, \tau_k(n)\}$, by exploiting the temporal correlation present in spectral speech data. Once underlying channel-specific models are parameterized, HMM-based decoding can be utilized to determine the minimum mean-square error (MMSE) estimate of the true binary SPP mask. The decoded SPP corresponding to the observed probability $\tau_k(n)$ is determined via the forward algorithm as [Rab89]

$$\hat{\tau}_k(n) = P(H_1(n) | \tau_k(1), \dots, \tau_k(n)) = \frac{\alpha_k^1(n)}{\alpha_k^0(n) + \alpha_k^1(n)} \quad (4.8)$$

where $\alpha_k^i(n)$ is the forward variable for channel k , corresponding to state i , at time index

n . Forward variables convey the probability of the current observation occupying state i , given past observations

$$\alpha_k^i(n) = P(H_i(n) | Y_k(1), \dots, Y_k(n)). \quad (4.9)$$

The forward variables $\alpha_k^i(n)$ and can be determined recursively as

$$\alpha_k^i(n) = \begin{cases} \left[\sum_{j=0}^1 a_{ji} \alpha_k^j(n-1) \right] b_k^i(\tau_k(n)), & \text{if } n > 1 \\ 1, & \text{else} \end{cases} \quad (4.10)$$

Thus, Equation 4.8 provides improved speech presence probabilities, given current and past standard SPPs.

4.3 Incorporating Future Observations

For applications in which slight delays are acceptable, improved SPPs can be determined by incorporating future observations via the forward-backward algorithm [Rab89]. In this case, similar to Eq. 4.8, the decoded speech presence probability is

$$\hat{\tau}_k(n) = P(H_1(n) | \tau_k(1), \dots, \tau_k(n + N_{LA})) = \frac{\alpha_k^1(n) \beta_k^1(n, 0)}{\alpha_k^1(n) \beta_k^1(n, 0) + \alpha_k^0(n) \beta_k^0(n, 0)} \quad (4.11)$$

where N_{LA} is the total number of look-ahead frames utilized. Here, $\beta_k^i(n, k)$, the backward variable, differs from traditional notation in that it is a function of two parameters. This is due to the generally time-sensitive nature of tasks such as speech enhancement,

which forces the recursive calculation of backward variables to be re-initialized for each time index n . The backward variable $\beta_k^i(n, k)$ conveys the probability of channel k occupying state i at time index $n + k$, during recursive calculations ultimately required for time index n

$$\beta_k^i(n, k) = P(H_i(n + k) | Y_k(n + k), \dots, Y_k(n + N_{LA})). \quad (4.12)$$

The backward variable $\beta_k^i(n, k)$ and can be determined recursively as

$$\beta_k^i(n, k) = \begin{cases} \sum_{j=0}^1 a_{01} \beta_k^j(n, k + 1) b_k^j(\tau_k(n + k + 1)), & \text{if } k < N_{LA} \\ 1, & \text{else} \end{cases} \quad (4.13)$$

In this way, future observations can be exploited to improve the estimation of standard SPPs.

4.4 Complexity Analysis

A well known downside to HMM-based processing is the induced computational load. This is especially problematic for speech applications, which can be delay-sensitive and/or resource constrained. However, due to the small size of the underlying model used during estimation of improved SPPs, the induced complexity is relatively small.

Table 4.1 provides operations required by the proposed algorithms for determining SPPs. The traditional method from [EM84], and the improved method from [Coh05] are included for reference. It can be observed from Table 4.1 that the additional number of operations required, as compared to [EM84], is relatively low, making it an attrac-

Table 4.1: Required operations for proposed SPPs: Numbers of operations are given per frame and per frequency channel. Note that the induced load of the fast Fourier transform (FFT) is not included, but is known to be of order $O(N_{ch} \log(N_{ch}))$.

SPP Method	+	\times	\div	cos	exp	log
Traditional SPPs[EM84]	4	6	3	0	1	0
Improved SPPs from [Coh05]	40	45	8	0	0	4
Proposed SPPs (Eq. 4.8)	12	20	5	2	1	0
Proposed SPPs (Eq. 4.11)	$12 + 2N_{LA}$	$20 + 8N_{LA}$	5	2	1	0

tive option for resource-constrained applications. Furthermore, methods proposed in this section induce a significantly smaller computational load than that of [Coh05]. In Table 4.1, number of operations is given per frame and per frequency channel. Note that the induced load of the fast Fourier transform (FFT) is not included, but is known to be of order $O(N_{ch} \log(N_{ch}))$, where N_{ch} denotes the number of channels used during spectral analysis.

4.5 Experimental Results

4.5.1 Accuracy of Improved SPPs

Figure 4.2 presents illustrative examples of SPP masks determined by various methods. Panel (a) provides the clean speech signal "She had your dark suit in greasy wash water all year" by a female speaker. Panel (b) shows the SPP mask determined according to [Coh05] from a corresponding signal corrupted by airport noise at 15 dB SNR. Panels (c) and (d) provide proposed SPP masks according to Eq. 4.8 and Eq. 4.11 ($N_{LA}=2$), respectively. As can be observed in panel (b), the algorithm proposed in [Coh05] results in a high false alarm rate. The proposed mask in panel (c) significantly reduces the false alarm rate, and manages to detect individual harmonics. Incorporating future observations in (d) results in a smoothed mask, wherein detected speech regions are presented in more

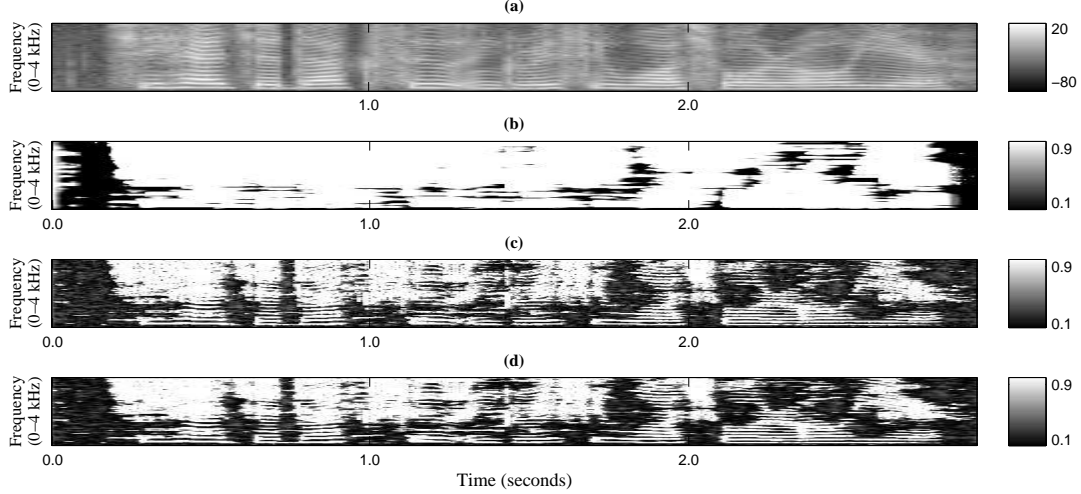


Figure 4.2: Illustrative examples of SPP masks determined by various methods: Panel (a) provides the clean speech signal "She had your dark suit in greasy wash water all year" spoken by a female. Panel (b) shows the SPP mask determined according to [Coh05] from the corresponding signal corrupted by airport noise at 15 dB SNR. Panels (c) and (d) provide proposed SPP masks according to Eq. 4.8 and Eq. 4.11 ($N_{LA}=2$), respectively.

salient segments.

To grade the accuracy of SPPs¹, pointwise Kullback-Leibler (KL) distances [CT91] are used between masks obtained from noisy speech and those obtained from corresponding "oracle" clean speech. The KL distance is suitable since it is commonly used to compare statistical distributions, and time- and frequency-specific SPPs are observations of individual pdfs. The mean pointwise KL distance between masks is given by

$$\bar{D}(\tau^{orc} || \hat{\tau}) = \frac{1}{N_t N_k} \sum_{n=1}^{N_t} \sum_{m=1}^{N_k} \tau_k^{orc}(n) \log \left(\frac{\tau_k^{orc}(n)}{\hat{\tau}_k(n)} \right), \quad (4.14)$$

where τ^{orc} refers to the oracle mask, and N_t denotes the length of the given sound file in frames. Note that oracle masks are determined by setting a hard energy threshold in

¹Application of algorithms described in 4 requires certain numerical parameters, which are included in Table 4.2

Table 4.2: Numerical parameters for proposed SPPs

Parameter	Enhancement	ASR	Description
N_m	257	257	number of channels during spectral analysis
a_m^{01}	0.05	0.20	HMM transitional probability, $H_m^0 \rightarrow H_m^1$
a_m^{10}	0.10	0.20	HMM transitional probability, $H_m^1 \rightarrow H_m^0$
κ^0	1.0	1.0	used by observation probability, $b_m^0(\tau_m(n))$
κ^1	10.0	7.0	used by observation probability, $b_m^1(\tau_m(n))$

Table 4.3: Mean pointwise Kullback-Leibler (KL) distances for SPP masks from oracle masks, in bits

SPP Method	SNR (dB)			
	15	10	5	0
from [Coh05]	0.21	0.33	0.52	0.77
Eq. 4.8	0.10	0.17	0.26	0.45
Eq. 4.11 ($N_{LA}=2$)	0.09	0.17	0.28	0.40

the spectral domain.

Table 4.3 provides mean pointwise KL distances for SPP masks from corresponding oracle masks, with results averaged across noise types. As reference, results for SPPs from [Coh05] are included. Note that similar to the proposed method in this paper, the work in [Coh05] exploits past observations. However, while [Coh05] combines current and past observations in a somewhat heuristic manner, the proposed method utilizes a statistical framework to find the MMSE estimate given the HMM framework. As can be concluded from Table 4.3, the proposed speech presence probabilities provide a significant increase in mask accuracy in terms of the pointwise KL distance. It should be noted that for certain cases, the KL distance increases with the inclusion of look-ahead frames. Using look-ahead frames tends to result in high SPPs occurring in salient segments due to a higher degree of HMM-based smoothing. It, in turn, provides lower missed detection rates, while increasing false alarms.

4.5.2 Speech Distortion and Noise Leakage

The performance of proposed SPP masks was tested by analyzing their ability to minimize acoustic noise leakage (NL) while maintaining low speech distortion. Similar to the experimental procedure in [GBM08], noise leakage is defined as the percentage of inactive time-frequency bins, which passes unsuppressed by SPP masks

$$NL = 100 \frac{\sum_{n=1}^{N_t} \sum_{m=1}^{N_k} \max [\hat{\tau}_k(n) - \tau_k^{orc}(n), 0] |Y_k(n)|^2}{\sum_{n=1}^{N_t} \sum_{m=1}^{N_k} \max [1 - \tau_k^{orc}(n), 0] |Y_k(n)|^2}. \quad (4.15)$$

Conversely, speech distortion (SD) is defined as the percentage of energy corresponding to active speech bins, which is distorted by SPP masks

$$SD = 100 \frac{\sum_{n=1}^{N_t} \sum_{m=1}^{N_k} \max [\tau_k^{orc}(n) - \hat{\tau}_k(n), 0] |Y_k(n)|^2}{\sum_{n=1}^{N_t} \sum_{m=1}^{N_k} \max [\tau_k^{orc}(n), 0] |Y_k(n)|^2}. \quad (4.16)$$

There exists a natural trade-off between NL and SD, i.e. as a greater percentage of acoustic noise is suppressed, more distortion to the underlying speech signal is generally expected.

Table 4.4 provides results for speech distortion and noise leakage for proposed SPP masks during soft-decision speech enhancement. The state-of-the-art method from [Coh05] is included as reference. Code for the previously discussed algorithms was obtained from [Coh]. Proposed SPPs (Eq. 4.8) are shown to provide low SD, while significantly decreasing the NL for most conditions. Integrating future observations in proposed SPPs (Eq. 4.11, $N_{LA}=1, 2$) generally results in a decrease in both SD and NL.

This chapter presents a framework for determining improved SPPs using HMM-based inference. By modeling spectro-temporal data as observations from channel-specific two-state models, and HMM-based decoding is applied to estimate true posterior probabilities. Improved SPP masks are shown to provide promising results with respect to various metrics.

Table 4.4: Speech distortion (SD) and noise leakage (NL) results for proposed SPP masks: Proposed techniques show low SD while providing significantly reduced NL, relative to [Coh05], for most noise conditions.

SPP Method	15 dB		10 dB		5 dB	
	SD (%)	NL (%)	SD (%)	NL (%)	SD (%)	NL (%)
from [Coh05]	0.2	69.9	0.7	56.4	2.3	43.6
Eq. 4.8	1.0	32.6	2.5	22.1	6.7	15.9
Eq. 4.11 ($N_{LA}=1$)	0.7	31.5	2.0	19.6	5.7	12.9
Eq. 4.11 ($N_{LA}=2$)	0.6	36.1	1.7	23.4	5.0	15.7

Part II

**Front-End Missing Feature Approaches
to Noise Robust ASR**

Part II explores front-end missing feature approaches to noise robust ASR. In Chapter 6.1, a statistical approach to Mel-domain mask estimation is discussed, which provides soft probabilistic metrics corresponding to individual spectral components. Next, to alternative MF spectral reconstruction solutions are introduced. Specifically, Chapter 6 proposes an HMM-based method of inferring missing data, whereas Chapter 7 explores the role of signal sparsity in spectral reconstruction. Finally, missing feature theory is extended to the task of packet loss concealment in Chapter 8.

It should be noted that experimental analysis of MF spectral reconstruction methods requires binary masks. Therefore, we utilize mask estimation from Chapter 6.1 during experimental procedures discussed in Chapters 6 and 7.

CHAPTER 5

A Statistical Approach to Mel-Domain Mask Estimation

This chapter proposes a novel approach to reliability mask estimation in the Mel-filtered spectral domain. Front-end missing feature approaches have been shown successful in the Mel-filtered domain [RS05]. The reason for this result is twofold: First, the Mel-filtering process significantly increases the level of inter-channel correlation, since it generally outputs spectra with unresolved harmonics. Many spectral reconstruction algorithms rely on inter-channel correlation during estimation of missing data. Second, Mel-filtering reduces the number of channels comprising spectra representations, thereby reducing the computational complexity of channel-based MF approaches. The framework proposed in this chapter operates on the Mel-filtered spectrum. However, it is designed to be robust to various frequency scales, such as the linear frequency scale and the Bark scale [Qua01].

As in Section 1.2.5, speech activity is modeled with a two-state Markov model, wherein state H_1 denotes time-frequency components corresponding to active speech, and state H_0 corresponds to time-frequency components comprised solely of noise. Following the Gaussian framework, power spectral coefficients of the observed signal are exponentially conditionally distributed for each state

$$\begin{aligned} p(|Y_k|^2|H_0) &= \frac{1}{\sigma_{N,k}^2} \exp\left(-\frac{|Y_k|^2}{\sigma_{N,k}^2}\right) \\ p(|Y_k|^2|H_1) &= \frac{1}{\sigma_{X,k}^2 + \sigma_{N,k}^2} \exp\left(-\frac{|Y_k|^2}{\sigma_{X,k}^2 + \sigma_{N,k}^2}\right). \end{aligned} \tag{5.1}$$

Note that the conditional distributions of Eq. 5.1 are special cases of the χ^2 distribution (see Appendix A.1.3) with $k=2$ degrees of freedom.

5.1 The Mel-Filtered Domain

As discussed in Section 1.3.1, automatic speech recognition (ASR) front ends generally include perceptually motivated frequency warping to emphasize discriminative information. For example, the Mel-filterbank is designed to approximate the human auditory system. As discussed in [YKO], the Mel-filterbank is defined in the spectral domain by the set of triangular filters $w_m(k)$, where m denotes Mel-channel index, and k denotes tap index, resulting in Mel-domain power-spectra expressed as

$$|\hat{Y}_m|^2 = \sum_{k=c_{m-1}}^{c_{m+1}} w_m(k) |Y_k|^2, \quad (5.2)$$

where c_m denotes the center frequency of the m^{th} -channel Mel-filter.

The distribution of the sum of weighted exponential random variables with distinct, possibly unequal, variances is often referred to as the generalized χ^2 distribution. Although a closed form expression for such distributions does not exist, several studies have proposed approximations which utilize Pearson curves, or moment matching to simpler χ^2 distributions [SS77].

As can be interpreted from Eq. 5.2, Mel-domain power spectral coefficients occur as weighted sums of exponential random variables with generally unequal variances, where weights correspond to Mel filter taps. The channel observation $|\hat{Y}_m|^2$ is modeled by a simpler χ^2 distribution with k_m degrees of freedom, where k_m is empirically determined using the ratio of noncentral moments

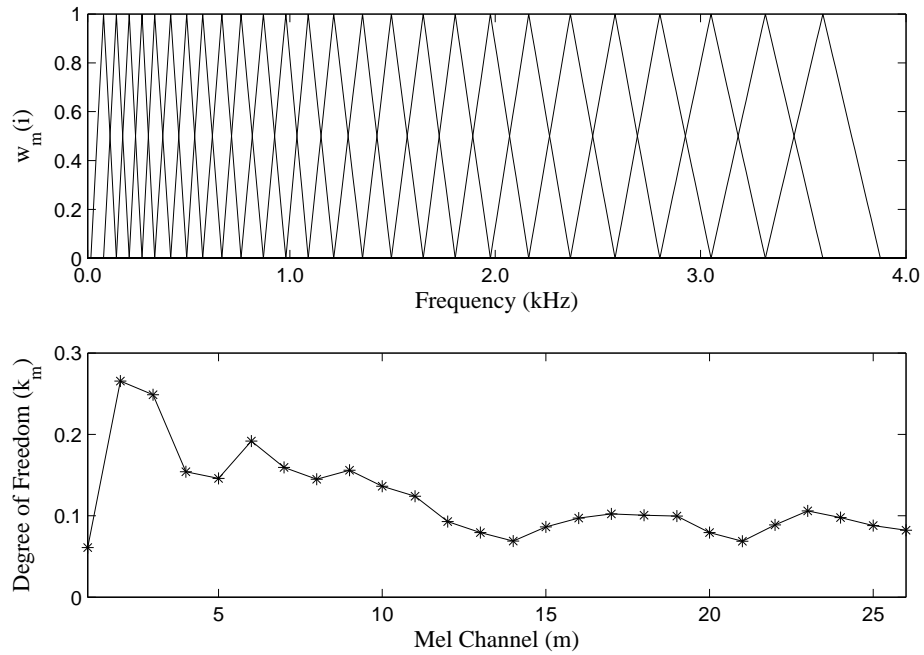


Figure 5.1: The Mel-filtering process and the effect on Mel-domain power spectra distributions: The top panel illustrates the Mel-filterbank, as defined in [YKO], for 26 channels. The bottom panel provides the resulting degrees of freedom, k_m , for Mel-channel power spectra χ^2 distributions, as determined empirically by Eq. 5.3.

$$k_m = \frac{2E \left[|\hat{Y}_m|^2 \right]^2}{E \left[|\hat{Y}_m|^4 \right]} = 2 \left(\frac{E \left[|\hat{Y}_m|^4 \right]}{E \left[|\hat{Y}_m|^2 \right]^2} - 1 \right)^{-1} \quad (5.3)$$

The top panel in Figure 5.1 illustrates the Mel-filterbank, as defined in [YKO], for 26 channels. The bottom panel provides the resulting degrees of freedom, k_m , for Mel-channel power spectra χ^2 distributions, as determined empirically by Eq. 5.3. Results in Figure 5.1 were obtained on clean training speech from the Aurora-2 database.

Furthermore, the observation $|\hat{Y}_k|^2$ is whitened by appropriate normalization with "weighted average" noise and signal variances corresponding to the m^{th} Mel channel, defined similarly to [Mar01]

$$\begin{aligned} \hat{\sigma}_{N,m}^2 &\triangleq \frac{\sum_{i=c_{m+1}}^{c_{m+1}} w_m(i) \sigma_{N,i}^2}{\sum_{i=c_{m+1}}^{c_{m+1}} w_k(i)} \\ \hat{\sigma}_{X,m}^2 &\triangleq \frac{\sum_{i=c_{m+1}}^{c_{m+1}} w_m(i) \sigma_{X,i}^2}{\sum_{i=c_{m+1}}^{c_{m+1}} w_k(i)} \end{aligned} \quad (5.4)$$

The conditional distributions of the m^{th} -channel Mel-domain power spectral coefficient is then approximated as

$$p \left(|\hat{Y}_m|^2 | H_0 \right) = \left(2^{k_m/2} \Gamma(k_m/2) \right)^{-1} \left(\frac{|\hat{Y}_m|^2}{\hat{\sigma}_{N,m}^2} \right)^{(k_m/2-1)} \exp \left(-\frac{|\hat{Y}_m|^2}{2\hat{\sigma}_{N,m}^2} \right) \quad (5.5)$$

$$p \left(|\hat{Y}_m|^2 | H_1 \right) = \left(2^{k_m/2} \Gamma(k_m/2) \right)^{-1} \left(\frac{|\hat{Y}_m|^2}{\hat{\sigma}_{N,m}^2 + \hat{\sigma}_{X,m}^2} \right)^{(k_m/2-1)} \exp \left(-\frac{|\hat{Y}_m|^2}{2(\hat{\sigma}_{N,m}^2 + \hat{\sigma}_{X,m}^2)} \right) \quad (5.6)$$

5.2 Mask Estimation

5.2.1 Soft-Decision Masks

This section utilizes the statistical framework developed in Section 1.2.5 to infer the probability of active speech in individual time-frequency bins. A Bayesian approach allows the posterior probability of active speech to be expressed as

$$P\left(H_1||\hat{Y}_m|^2\right) = \frac{\Lambda_m}{1 + \Lambda_m}, \quad (5.7)$$

where the Generalized Likelihood Ratio (GLR) is defined as

$$\Lambda_m = \eta_m \frac{p\left(|\hat{Y}_m|^2|H_1\right)}{p\left(|\hat{Y}_m|^2|H_0\right)}, \quad (5.8)$$

and η_m denotes the ratio of steady-state probabilities. Substitution of Eq. 5.5 into Eq. 5.8 leads to

$$\Lambda_m(n) = \eta_m \left(\frac{\hat{\sigma}_{N,m}^2}{\hat{\sigma}_{N,m}^2 + \hat{\sigma}_{X,m}^2} \right)^{(k_m/2-1)} \exp\left(\frac{|\hat{Y}_m|^2 \hat{\sigma}_{X,m}^2}{2\hat{\sigma}_{N,m}^2 (\hat{\sigma}_{N,m}^2 + \hat{\sigma}_{X,m}^2)} \right) \quad (5.9)$$

Analogous to the linear frequency domain parameters presented in [MM80], Mel-domain *a priori* and *a posteriori* SNRs are defined as

$$\hat{\xi}_m(n) \triangleq \frac{\hat{\sigma}_{X,m}^2}{\hat{\sigma}_{N,m}^2}, \quad \hat{\gamma}_m(n) \triangleq \frac{|\hat{Y}_m|^2}{\hat{\sigma}_{N,m}^2}. \quad (5.10)$$

Since the term $\sigma_{X,m}^2$ is hidden, the *a priori* SNR is approximated similarly to the maximum likelihood approach presented in [EM84]

$$\hat{\xi}_m(n) \approx \max\{\bar{\gamma}_m(n) - 1, 0\}, \quad (5.11)$$

where the smoothed *a posteriori* SNR is determined recursively as

$$\bar{\gamma}_m(n) = \delta \bar{\gamma}_m(n-1) + (1 - \delta) \hat{\gamma}_m(n), \quad (5.12)$$

and where $0 \ll \delta < 1$. Eq. 5.9 then reduces to

$$\Lambda_m(n) = \eta_m \left(\frac{1}{1 + \hat{\xi}_m(n)} \right)^{(k_m/2-1)} \exp \left(\frac{\hat{\gamma}_m(n) \hat{\xi}_m(n)}{2(1 + \hat{\xi}_m(n))} \right) \quad (5.13)$$

Substitution of Eq. 5.13 into Eq. 5.7 reveals time- and channel-specific probabilities of active speech, which comprise the Mel-domain speech presence probability mask.

As opposed to previous solutions to soft-decision mask estimation for MF-based ASR, such as [GBC01], which uses a tunable sigmoid function to map SNR-related measures to the range $[0, 1]$, the proposed method provides a statistical approach which offers intuitive probabilistic mask values. Additionally, the proposed mask estimation method relies on a simple training process which requires only clean speech. Other similar work, such as [KS06] and [SRS04], includes extensive training to determine the empirical distribution of classifiers in various noise types and levels. The mask estimation technique in [Sh04] requires training of phonetic class-dependent vector quantizer codebooks.

5.2.2 HMM-Based Decoding

Section 5.2.1 proposes speech presence probabilities masks comprised of posterior probabilities conditioned on single Mel-domain power spectrum observations $|\hat{Y}_m|^2$. In this section, HMM-based decoding is applied to exploit the well-known temporal correlation

property of speech spectra. Following [Rab89], the forward and backward variables are defined recursively as

$$\alpha_m^i(n) = P\left(H_i \mid |\hat{Y}_m(1)|^2, \dots, |\hat{Y}_m(n)|^2\right) = \left[\sum_{j=0}^1 a_{ji} \alpha_m^j(n-1) \right] p\left(|\hat{Y}_m(n)|^2 \mid H_i\right) \quad (5.14)$$

$$\beta_m^i(n) = \quad (5.15)$$

$$P\left(H_i \mid |\hat{Y}_m(n)|^2, \dots, |\hat{Y}_m(n+N_{la})|^2\right) = \sum_{j=0}^1 a_{ji} \beta_m^j(n+1) p\left(|\hat{Y}_m(n+1)|^2 \mid H_j\right)$$

where a_{ij} represents the transitional probability from state H_i to state H_j , and N_{la} denotes the number of look-ahead frames used during backward estimation. Note that with the assumption of a 2-state Markov model, the ratio of steady-state probabilities can be obtained from transitional statistics as $\eta_m = a_{01}/a_{10}$.

The ratio of forward variables is defined as

$$\Phi_m(n) = \frac{\alpha_m^1(n)}{\alpha_m^0(n)} = \frac{a_{01} + a_{11}\Phi_m(n-1)}{a_{00} + a_{10}\Phi_m(n-1)} \Lambda_m(n), \quad (5.16)$$

and the ratio of backward variables as

$$\Psi_m(n) = \frac{\beta_m^1(n)}{\beta_m^0(n)} = \frac{a_{10} + a_{11}\Psi_m(n+1) \Lambda_m(n+1)}{a_{00} + a_{01}\Psi_m(n+1) \Lambda_m(n+1)}. \quad (5.17)$$

Since Eqs. 5.16 and 5.17 are expressed recursively, boundary conditions are provided as

$$\Phi_m(1) = \frac{a_{01}}{a_{01} + a_{10}} \quad (5.18)$$

$$\Psi_m(n + N_{la}) = \frac{a_{01}}{a_{01} + a_{10}} \quad (5.19)$$

Note that a similar expression to Eq. 5.16 was presented in [SKS99], but was applied on a global, channel-independent basis, for the task of voice activity detection. Using Eq. 5.8 and Eqs. 5.14-5.17, the decoded GLR is defined as

$$\tilde{\Lambda}_m(n) = \frac{\alpha_m^1(n) \beta_m^1(n)}{\alpha_m^0(n) \beta_m^0(n)} = \Phi_m(n) \Psi_m(n). \quad (5.20)$$

Substitution of Eq. 5.20 into Eq. 5.7 reveals a soft-decision speech presence uncertainty mask which exploits temporal correlation of speech spectra.

5.2.3 Binary Masks

Although soft-decision masks are generally useful in noise robust speech processing, the majority of missing feature data inference techniques for noise robust ASR, such as [BA09] and [BA10a], require binary masks which differentiate between reliable and unreliable Mel-domain components. Probabilistic values derived in previous sections can be mapped to binary values via hard-thresholding with a parameter (ς). Figure 5.2 provides illustrative examples of Mel-domain mask estimation. Panel **a** shows the clean utterance "nine one nine six nine five one," from the Aurora-2 database. Panels **b** and **c** provide the proposed soft masks, without and with HMM-based decoding, respectively, for the corresponding speech signal degraded by vehicle noise at 5 dB SNR. Panel **d** illustrates the proposed binary mask obtained by hard thresholding.

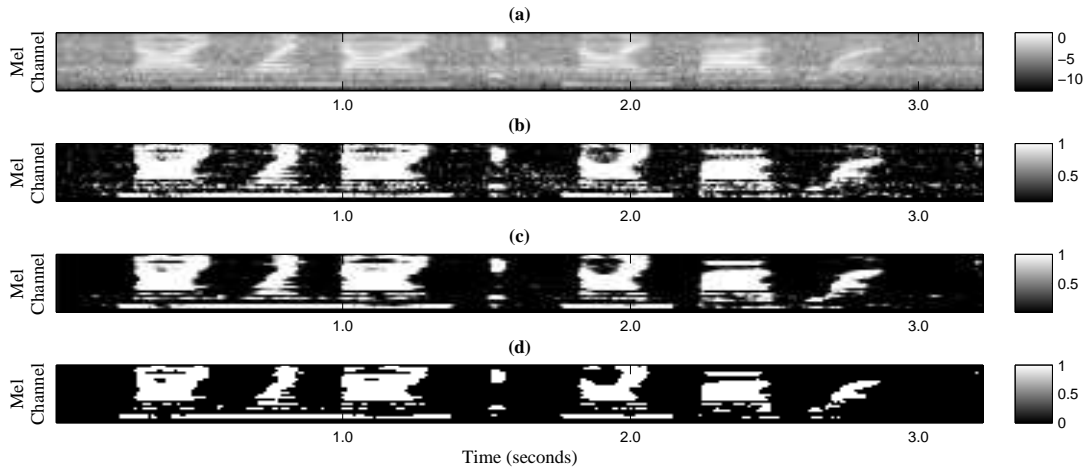


Figure 5.2: Examples of Mel-domain mask estimation: Panel **a** shows the clean utterance “nine one nine six nine five one,” from the Aurora-2 database. Panels **b** and **c** provide the proposed soft masks, without and with HMM-based decoding, respectively, for the corresponding speech signal degraded by vehicle noise at 5 dB SNR. Panel **d** illustrates the proposed binary mask obtained by hard thresholding.

5.3 Experimental Results

To assess the accuracy of the proposed Mel-domain mask estimation framework, it is applied to MF-based ASR. As an illustrative example the spectral reconstruction technique from Chapter 7 is used in combination. The overall ASR system is applied to the Aurora-2 database, which consists of connected digit utterances and contains 8 types of non-stationary noise at SNR levels within the range of 20 to -5 dB. The baseline recognizer is constructed using HTK code [YKO] and uses 16-state, 3-mixture word models. Feature extraction includes 12-dimensional Mel-frequency cepstral coefficients (MFCCs) with log-energy, along with first and second derivatives [RH93]. ASR results include cepstral domain mean normalization (CMN) [VL98].

The baseline mask estimation technique uses a Mel-domain version of the mask estimation technique proposed in [VGC99], which determines reliable time-frequency components as the intersection of two criteria

Table 5.1: Parameters utilized during mask estimation

Parameter	Value	Description
δ	0.25	Forgetting factor from Eq. 5.12
a_{01}	0.15	Transitional probability, $H_0 \rightarrow H_1$
a_{10}	0.35	Transitional probability, $H_1 \rightarrow H_0$
ς	0.75	Thresholding parameter

Table 5.2: Word-accuracy results for missing feature ASR using various mask estimation techniques, and applying the compressive sensing (CS)-based spectral reconstruction method from Chapter 7

SNR (dB)	20	15	10	5	0	Ave.
BL	95.8	89.6	70.4	34.6	5.2	59.1
SS-Based [VGC99]	97.2	95.0	88.6	72.1	39.6	78.5
Proposed	97.6	95.9	91.0	79.4	53.4	83.5
Oracle	97.6	96.9	95.1	91.0	80.9	92.3

(i) The Negative Energy Criterion: $|\hat{Y}_m| - |\hat{N}_m| > 0$

(ii) The SNR Criterion: $\left(|\hat{Y}_m| - |\hat{N}_m|\right)^2 > 1/2|\hat{Y}_m|^2$

As opposed to other prior mask estimation methods, [VGC99] provides a fair baseline, since neither it nor the proposed technique require noisy speech during training.

Table 5.2 provides word-accuracy results for MF-based ASR using proposed mask estimation in combination with the spectral reconstruction technique from Chapter 7.1. Results for spectral subtraction (SS)-based mask estimation from [VGC99] are included as reference. Additionally, results for the baseline system (BL) without spectral reconstruction are included. As can be observed, the proposed mask estimation method provides significant improvements in word-accuracy relative to that of [VGC99].

This chapter presents a statistical approach to Mel-domain mask estimation in which reliability measures are derived as conditional probabilities of active speech using a Bayesian

approach. Mel-domain power spectra are modeled as χ^2 processes with empirically-determined degrees of freedom. The proposed mask estimation algorithm is applied to the compressive sensing-based MF spectral reconstruction technique from Chapter 7, and is shown to outperform the baseline method from [VGC99] in terms of word-accuracy.

CHAPTER 6

HMM-Based Reconstruction of Missing Features

6.1 Noise Robust Feature Extraction: Log-Spectral Flooring

The aim of designing noise robust front-end features for ASR is to maximize discriminative spectral information while minimizing variability due to noise [ZA03]. It has been widely reported that discriminative speech information tends to lie in high amplitude spectral peaks, whereas noise tends to lie in spectral valleys. Such observations have been motivated by human perception [SA97] and shown quantitatively [ZA03]. Resulting efforts to overcome ASR performance degradation due to noise include spectral flooring [ZA03] and peak isolation [SA97]. In this section, a simple algorithm is proposed for front-end feature extraction which aims to minimize feature variability due to noise by flooring observed data in the log-spectral domain. The technique is utilized during MF spectral reconstruction throughout this dissertation.

Let \tilde{R}_k represent the log Mel-filtered spectral amplitude of an observed speech signal, after liftering. The logarithmic function is applied during feature extraction as it is similar to what the human auditory system does, and effectively compands the energy scale. This serves to emphasize discriminative speech patterns present in the spectral envelope, which are considered important for recognition [SA97]. However, the logarithm results in an unbounded lower limit, leading to an increased dynamic range for spectral valleys. This may cause confusion during recognition since spectral valleys typically contain little discriminative speech energy. To reduce the confusability introduced by the logarithmic function on small spectral values, half-wave rectification is performed relative to a prede-

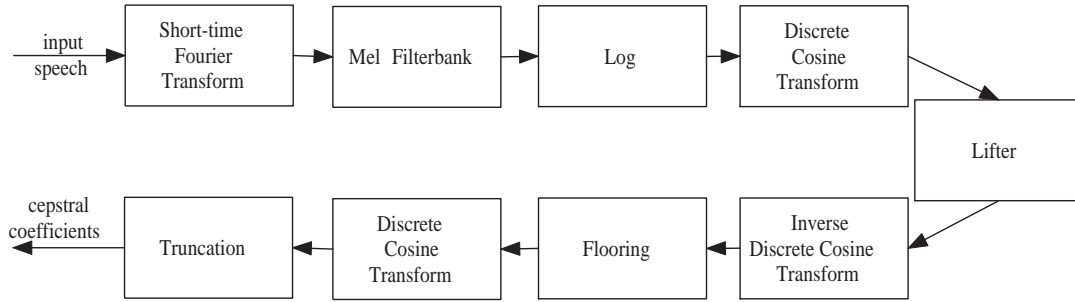


Figure 6.1: Overview of the log-spectral flooring feature extraction process

terminated threshold. Log-spectral flooring (LS-FLR) is thus defined as

$$\tilde{R}_k = \begin{cases} \tilde{R}_k, & \text{if } \tilde{R}_k \geq \alpha_{fl} \\ \alpha_{fl}, & \text{otherwise} \end{cases} \quad (6.1)$$

where α_{fl} determines the flooring threshold. In this study, the flooring parameter was empirically optimized to $\alpha_{fl} = 0.0$ dB. It should be noted that the deterministic threshold α_{fl} may need to be tuned to match the dynamic range of the input signal. Figure 6.1 provides a graphical overview of the log-spectral flooring feature extraction process. It is interesting to note the difference between Figure 6.1 and the standard front-end illustrated in Figure 1.10.

Figure 6.2 provides an illustrative example of log-spectral flooring. The top panel shows the clean and noisy versions of liftered log-spectra corresponding to an active frame of speech ($/\varepsilon/$ from "seven"). The bottom panel shows the same spectra after log-spectral flooring. The noisy speech is degraded by vehicular noise at 10 dB SNR. LS-FLR reduces variability between clean and noisy spectral features, which is visible in lower Mel channels. Additionally, flooring reduces the dynamic range of log-spectra by instituting a lower bound below which little discriminative speech energy typically lies.

Figure 6.3 provides an alternate illustrative perspective of log-spectral flooring. The top panel shows the Mel-filtered spectrogram of the utterance "one two three seven seven four three," from the Aurora-2 database, degraded by car noise at 10 dB. The middle panel

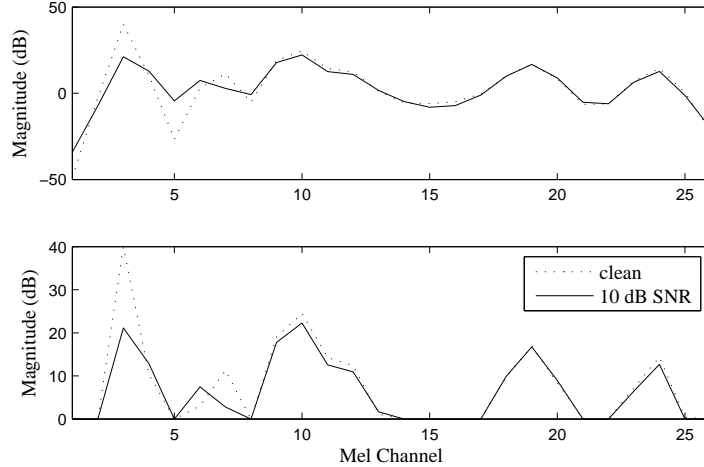


Figure 6.2: An illustrative example of log-spectral flooring: The top panel shows the clean and noisy versions of liftered log-spectra corresponding to an active frame of speech ($/\varepsilon/$ from "seven"). The bottom panels shows the same spectra after log-spectral flooring. The noisy speech was degraded by car noise at 10 dB SNR.

shows the spectrogram after liftering. The bottom panel shows the spectrogram after log-spectral flooring. It can be observed that log-spectral flooring emphasizes discriminative spectral information, and minimizes the apparent presence of noise.

The proposed LS-FLR algorithm is applied to the Aurora-2 database, using the experimental procedure described in Chapter , and Table 6.1 provides word-accuracy results. For comparison, traditional spectral flooring [ZA03] is included. It can be observed in Table 6.1 that flooring in the log-spectral domain provides significant performance improvements relative to traditional spectral flooring, and thus serves as an efficient and effective front end noise robust feature.

6.2 The Role of HMMs in Spectral Reconstruction

This chapter explores HMM-based reconstruction of unreliable spectro-temporal data. The framework presented here exploits correlation along the time and/or frequency axes to infer degraded components. Furthermore, a method by which to downsample HMMs

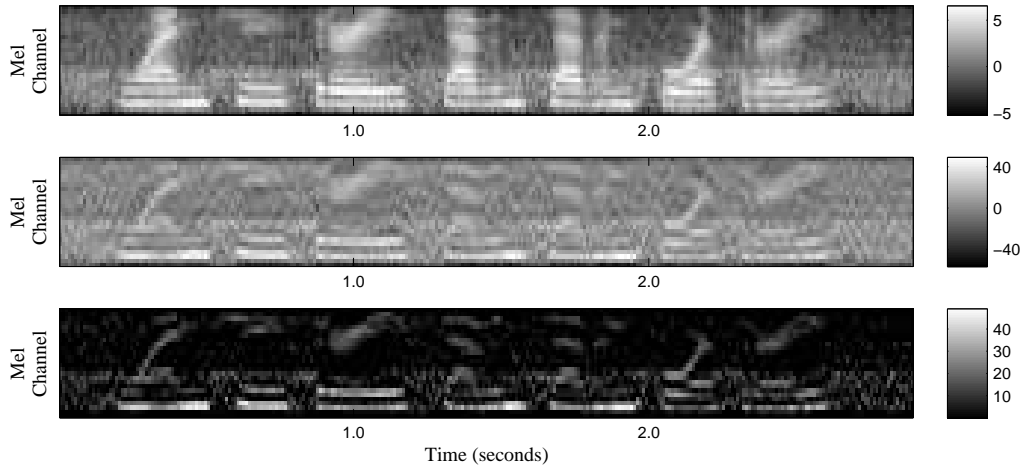


Figure 6.3: An illustrative example of log-spectral flooring: The top panel shows the Mel-filtered spectrogram of the utterance "one two three seven seven four three," from the Aurora-2 database, degraded by car noise at 10 dB SNR. The middle panel shows the spectrogram after liftering. The bottom panel shows the spectrogram after log-spectral flooring.

Table 6.1: Word-Accuracies for the Aurora-2 database using log-spectral flooring (LS-FLR). Spectral flooring (S-FLR) is included for comparison.

SNR (dB)	20	15	10	5	0	Ave
none	95.8	89.6	70.4	34.6	5.2	59.1
S-FLR	96.8	93.3	82.3	55.5	14.2	68.4
LS-FLR	97.6	95.2	88.2	69.6	34.2	77.0

is proposed, leading to a significantly reduced computational load with negligible performance degradation.

A common method to model data trajectories is with an HMM ([PSS01],[PSP02],[BA09]). Due to the discrete nature of HMM states, features must be quantized, at least implicitly, prior to the estimation process. Note that when used in this study, the term quantizer is used slightly differently than in the traditional sense, such as in source coding. The aim of quantization in our work is not data compression, but rather to designate to each decoded feature element a corresponding HMM state.

Let the quantizer of channel m be represented by the set of scalar-valued centroids $\{c_m^1, c_m^2, \dots, c_m^N\}$. Channel-specific quantizers can be designed based on empirical data using clustering techniques such as the K-means algorithm [Har75]. In this study, quantizers were trained on clean data from the Aurora-2 database.

Let the quantization of the clean spectro-temporal feature $A_m(n)$ be performed by minimizing the Euclidean distance

$$q_m(A_m(n)) = c_m^i, \text{ where } i = \arg \min_j \|A_m(n) - c_m^j\|^2. \quad (6.2)$$

Furthermore, let the set $\{z_m^{0,1}, z_m^{1,2}, \dots, z_m^{N,N+1}\}$ be defined as boundaries between centroids such that the region in feature space confined by $z_m^{i-1,i}$ and $z_m^{i,i+1}$ can be interpreted as the cell corresponding to c_m^i . Boundary values are determined as

$$z_m^{i,i+1} = \begin{cases} \frac{1}{2}(c_m^i + c_m^{i+1}), & \text{if } 0 < i < N, \\ 0, & \text{if } i = 0, \\ \infty, & \text{if } i = N \end{cases}. \quad (6.3)$$

In order to apply estimation methods, separate HMMs are constructed for each channel-specific quantizer to model feature trajectories. Let the HMM applied to the output signal from channel m be referred to as $\Xi_m = (\mathbf{A}_m, \mathbf{B}_m, \boldsymbol{\pi}_m)$, where \mathbf{A}_m provides tran-

sitional statistics, \mathbf{B}_m provides observation statistics, and $\boldsymbol{\pi}_m$ provides steady-state statistics [Rab89]. The steady-state probabilities of Ξ_m can be determined empirically from training data as

$$\pi_i^m = \frac{\text{no. of samples quantized to centroid } c_m^i(n)}{\text{total no. of samples}}. \quad (6.4)$$

The transitional probabilities of Ξ_m can similarly be determined from training data as

$$a_{ij}^m = \frac{\text{no. of samples transitioning from } c_m^i(n) \text{ to } c_m^j(n+1)}{\text{no. of samples quantized to } c_m^i(n)}. \quad (6.5)$$

In this study, steady-state and transitional statistics were obtained from Aurora-2 clean training data.

Figure 6.4 shows example graphical versions of transitional probability matrices obtained empirically from the clean training data in the Aurora-2 database [Pea00]. For illustrative purposes, log-probabilities are shown. As can be observed, there exist strong diagonal trends in the transitional probability matrices. These diagonal trends correspond to low-entropy signals, and such statistical patterns further motivate the use of transitional probabilities during the estimation process.

The observation statistic $b_i^m(R_m(n))$ represents the probability that the underlying clean speech component $A_m(n)$, corresponding to centroid c_m^i , is observed as the noisy feature $R_m(n)$

$$b_i^m(R_m(n)) = p(R_m(n) | c_m^i). \quad (6.6)$$

Steady-state and transitional statistics are described by discrete distributions due to the discrete configuration of underlying signal models. Observation statistics, on the other hand, are continuous distributions, since they incorporate the reliability of noise magnitude estimation.

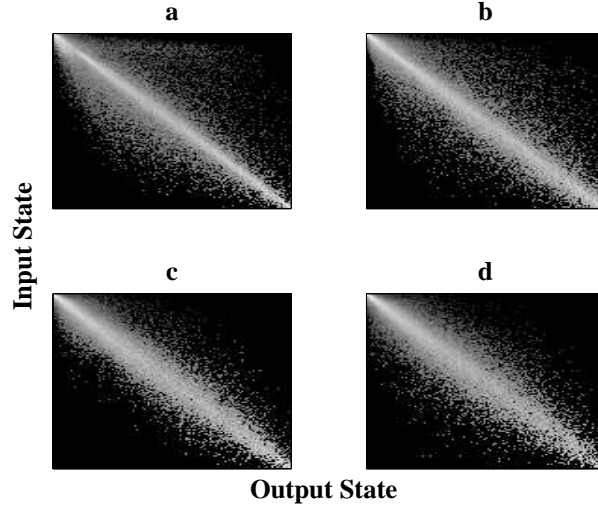


Figure 6.4: Example transitional probability matrices, \mathbf{A}_m : For illustrative purposes, log-probabilities are shown. The y-axes correspond to the input state and the x-axes correspond to the output state. Panels **a** through **d** refer to Mel-filtered Short-Time Fourier Transform channels with center frequencies 313 Hz, 734 Hz, 1328 Hz, and 2188 Hz, respectively.

The proposed algorithm builds upon an initial estimate of the clean underlying speech spectrum. While more complex estimators can be utilized, here, a simple initial spectral approximation which utilizes magnitude spectral subtraction (MSS) [Bol79]

$$\hat{A}_m(n) = R_m(n) - \hat{D}_m(R_m(n)), \quad (6.7)$$

where $\hat{D}_m(R_m(n))$ represents a local estimate of the noise spectral magnitude. It is assumed that the estimate $\hat{D}_m(R_m(n))$ is obtained with a certain level of accuracy, such that the actual hidden noise process, $D_m(n)$, can be expressed as

$$D_m(n) = \hat{D}_m(R_m(n)) + \varepsilon(n), \quad (6.8)$$

where $\varepsilon_m(n)$ is the estimation error. $\varepsilon(n)$ is assumed to be a stationary zero-mean Gaussian process with standard deviation σ_ε . By assuming $\varepsilon(n)$ to be zero-mean, it implies

that the noise estimation technique utilized does not produce bias. However, if noise estimation is known to produce bias, this can be addressed by shifting the mean of $\varepsilon(n)$.

Note that σ_ε is related to the accuracy of the noise estimation process. For example, if $\hat{D}_m(R_m(n))$ can be determined exactly, then $\sigma_\varepsilon=0$; conversely as the noise estimation process becomes less accurate, σ_ε increases. It therefore follows intuitively that the standard deviation of the noise estimation process should be related to the magnitude of the estimated noise. In this study, a rough approximation is used

$$\sigma_\varepsilon = \kappa \hat{D}_m(R_m(n)), \quad (6.9)$$

where κ was empirically set to 0.6.

Assuming a Gaussian distribution for the estimation error of the noise spectral magnitude, the observation probability distribution is given by

$$b_i^m(R_m(n)) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \int_{z_m^{i-1,i}}^{z_m^{i,i+1}} e^{-\frac{(\tau - \hat{D}_m(R_m(n)))^2}{2\sigma_\varepsilon^2}} d\tau, \quad (6.10)$$

where σ_ε is approximated by Equation 6.9.

It is important to note that the methods described for obtaining the relationship in Equation 6.9 should not be considered the matched-case or stereo-training scenario, as in [DAP00], which requires extensive corresponding clean and noisy speech databases. Instead, Equation 6.9 is a rough approximation involving a single parameter which is used to adapt the system to large variations in component-specific input SNR. More advanced relationships between σ_ε and $\hat{D}_m(R_m(n))$ can be developed, and may be expected to result in more accurate observation probabilities.

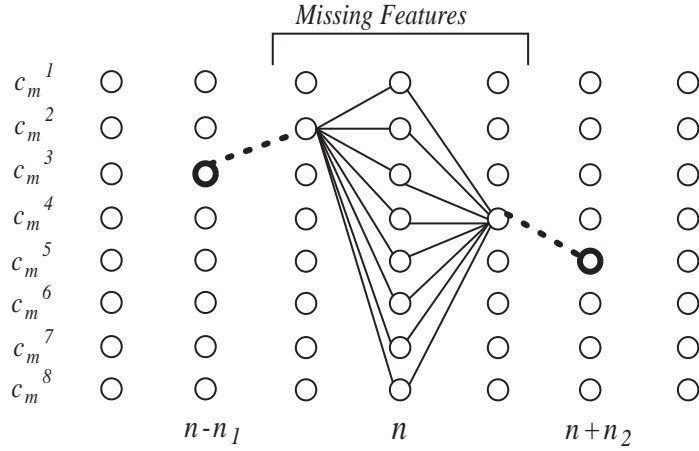


Figure 6.5: An illustrative example of the HMM-based missing feature problem formulation: In this case, the quantizer is comprised of $N=8$ centroids. A series of 3 features are missing. In this figure, the feature at time index n is to be estimated, so that $n_1=2$ and $n_2=2$. Note that in general, n_1 and n_2 are not equal.

6.3 HMM-Based Estimation Methods

Once underlying signal sources have been modeled by the set $\{\Xi_1, \dots, \Xi_M\}$, unreliable spectrographic data can be approximated by means of HMM-based estimation. This section presents HMM-based spectral reconstruction methods which exploit intra-channel correlation, inter-channel correlation, or a combination of both.

6.3.1 Utilizing Correlation Across Time

The forward-backward algorithm provides an accurate algorithm for estimating missing or ambiguous data within an HMM framework [PSS01],[PSP02]. The FB algorithm determines the MMSE spectral estimate $\hat{A}_m(n)$ given the first and last reliable features at temporal indices $n - n_1$ and $n + n_2$, and given the series of observations $R_m(n - n_1), \dots, R_m(n + n_2)$. Figure 6.5 provides an illustrative example of the missing feature problem formulation.

The estimate of component $A_m(n)$, using the FB algorithm and exploiting intra-channel

correlation, is determined as [Rab89]

$$\hat{A}_m(n) = \sum_{i=1}^N c_m^i \gamma_{T,m}^i(n), \quad (6.11)$$

where

$$\gamma_{T,m}^i(n) = \frac{\alpha_m^i(n) \beta_m^i(n)}{\sum_{j=1}^N \alpha_m^j(n) \beta_m^j(n)}. \quad (6.12)$$

The set of values $\gamma_{T,m}^i$ represents the conditional distribution of $q_m(A_m(n))$, conditioned on the first and last reliable features, as well as on past and future noisy observations

$$\gamma_{T,m}^i(n) = P(c_m^i | R_m(n - n_1), \dots, R_m(n + n_2)). \quad (6.13)$$

Note that the subscript T denotes the application of HMM-based decoding across time. The values $\alpha_m^i(n)$ and $\beta_m^i(n)$, known as the forward and backward variables, respectively, can be determined recursively as

$$\begin{aligned} \alpha_m^i(n) &= \left[\sum_{j=1}^N a_{ji}^m \alpha_m^j(n-1) \right] b_i^m(R_m(n)), \\ \beta_m^i(n) &= \sum_{j=1}^N a_{ij}^m \beta_m^j(n+1) b_j^m(R_m(n+1)). \end{aligned} \quad (6.14)$$

The forward variable conveys the probability of a given centroid conditioned on past and present observations, whereas the backward variable conveys the probability of a given centroid conditioned on future observations

$$\begin{aligned}\alpha_m^i(n) &= P(c_m^i | x_{n-n_1}, \dots, x_n) \\ \beta_m^i(n) &= P(c_m^i | x_{n+1}, \dots, x_{n+n_2}).\end{aligned}\tag{6.15}$$

The forward and backward variables at the last and first reliable feature vectors are known as

$$\alpha_m^i(n - n_1) = \begin{cases} 1, & \text{if } q_m(\hat{A}_m(n - n_1)) = c_m^i \\ 0, & \text{else} \end{cases}\tag{6.16}$$

and

$$\beta_m^i(n + n_2) = \begin{cases} 1, & \text{if } q_m(\hat{A}_m(n + n_2)) = c_m^i \\ 0, & \text{else} \end{cases}.\tag{6.17}$$

The HMM-based spectral reconstruction method utilizing intra-channel (across-time) correlation is referred to as **FB_T**.

6.3.2 Utilizing Correlation Across Frequency Channels

In many missing data scenarios, such as speech communication and remote speech recognition, unreliable data is a result of corrupt bits or dropped packets ([PSS01], [PSP02], [BA09]), and thus utilizing correlation along the frequency axis is not applicable. In the current study, however, this is not the case, and the strong inter-channel correlation present within speech can be exploited during spectrogram reconstruction. It is interesting to note that [RSS04] previously utilized inter-channel correlation by employing a full covariance Gaussian mixture model during imputation of missing spectro-temporal components of speech. When performing estimation along the frequency axis, the stationarity assumed

by the transitional statistics of a_{ij}^m does not hold. It is therefore necessary to introduce the probabilities

$$P_{k:l}^{i:j} = \begin{cases} P(q_k(A_k(n)) = c_k^i | q_l(A_l(n)) = c_l^j), & \text{for } |k - l| = 1, \\ 0, & \text{else} \end{cases} \quad (6.18)$$

which define the inter-channel transitional statistics of spectrographic data. That is, $P_{k:l}^{i:j}$ represents the probability that the spectrographic feature in channel k corresponds to centroid i , given that the spectrographic feature in channel l is quantized to centroid j . Note that since HMM-based processing assumes a 1st order signal model, then $|k - l|=1$ must hold. The probability $P_{k:l}^{i:j}$ can be determined from training data as

$$P_{k:l}^{i:j} = \frac{\text{no. of samples transitioning from } c_k^i(n) \text{ to } c_l^j(n)}{\text{no. of samples quantized to } c_k^i(n)}. \quad (6.19)$$

Modified forward and backward variables can be expressed for the estimation of missing data along the frequency axis

$$\delta_m^i(n) = \left[\sum_{j=1}^N P_{m:m-1}^{i:j} \delta_{m-1}^j(n) \right] b_i^m(R_m(n)), \quad (6.20)$$

$$\epsilon_m^i(n) = \sum_{j=1}^N P_{m:m+1}^{j:i} \epsilon_{m+1}^j(n) b_j^{m+1}(R_{m+1}(n)). \quad (6.21)$$

Note that $\delta_m^i(n)$ and $\epsilon_m^i(n)$ are determined similarly to the forward and backward variables expressed in Eq. (6.14), with the exception that the stationary transitional probability a_{ij}^m is replaced by $P_{k:l}^{i:j}$. $\delta_m^i(n)$ and $\epsilon_m^i(n)$ each represents a distribution of $q_m(A_m)$ conditioned on different subsets of inter-channel observations

$$\begin{aligned}\delta_m^i(n) &= P(c_m^i | R_{m-m_1}(n), \dots, R_m(n)) \\ \epsilon_m^i(n) &= P(c_m^i | R_{m+1}(n), \dots, R_{m+m_2}(n)).\end{aligned}\quad (6.22)$$

Here, $m - m_1$ and $m + m_2$ denote the channel indices of the first and last reliable components within the current feature vector. In this application, referred to as \mathbf{FB}_F , $\gamma_{F,m}^i(n)$ represents the distribution of $q_m(A_m(n))$ conditioned on reliable components and noisy observations from the feature vector at time index n

$$\gamma_{F,m}^i(n) = P(c_m^i | R_{m-m_1}(n), \dots, R_{m+m_2}(n)). \quad (6.23)$$

The corresponding expression for determining $\gamma_{F,m}^i(n)$ is given by

$$\gamma_{F,m}^i(n) = \frac{\delta_m^i(n) \epsilon_m^i(n)}{\sum_{j=1}^N \delta_m^j(n) \epsilon_m^j(n)}. \quad (6.24)$$

The forward and backward variables for inter-channel estimation are known at the last and first reliable spectral components as

$$\delta_{m-m_1}^i(n) = \begin{cases} 1, & \text{if } q_{m-m_1}(s_{m-m_1}(n)) = c_{m-m_1}^i \\ 0, & \text{else} \end{cases} \quad (6.25)$$

and

$$\epsilon_{m+m_2}^i(n) = \begin{cases} 1, & \text{if } q_{m+m_2}(s_{m+m_2}(n)) = c_{m+m_2}^i \\ 0, & \text{else} \end{cases}. \quad (6.26)$$

6.3.3 Utilizing Correlation Across Time and Across Frequency Channels

Additionally, estimation of unreliable spectral components can be performed as to utilize statistics along both the time and frequency axes, referred to as \mathbf{FB}_{2D} . In this scenario, $\gamma_{2D,m}^i(n)$ is the distribution of $R_m(n)$ conditioned on data contained in both channel i and feature vector $\mathbf{f}(n)$

$$\begin{aligned} \gamma_{2D,m}^i(n) = & P(c_m^i | R_m(n - n_1), \dots, R_m(n + n_2)) \\ & \times P(c_m^i | R_{m-m_1}(n), \dots, R_{m+m_2}(n)) \end{aligned} \quad (6.27)$$

Equation 6.27 leads to the conditional probabilities $\gamma_{2D,m}^i(n)$ being expressed as

$$\gamma_{2D,m}^i(n) = \gamma_{T,m}^i(n) \gamma_{F,m}^i(n) = \frac{\alpha_m^i(n) \beta_m^i(n) \delta_m^i(n) \epsilon_m^i(n)}{\sum_{j=1}^N \alpha_m^j(n) \beta_m^j(n) \sum_{k=1}^N \delta_m^k(n) \epsilon_m^k(n)}. \quad (6.28)$$

For the conditional distributions given by Equations 6.24 and 6.28, the corresponding estimates $\hat{A}_m(n)$ are determined similarly to Equation 6.11.

6.4 Efficient Approximation of HMM-Based Estimation Techniques

6.4.1 Downsampling of Statistical Models

A well-known downside to HMM-based processing is the induced computational load. For large codebooks (large N), the required complexity may cause HMM-based algorithms to be infeasible for resource constrained or delay-sensitive applications. A framework is discussed for efficient HMM-based missing feature estimation based on downsampling of underlying statistical models, previously presented in [BA09] and [BA08a].

Instead of using the original quantizer centroids to construct signal models, quantizers with less resolution are used to build statistical models. A tree-structure mapping of centroids is implemented to allow downsampling of the discrete HMMs by factors of 2, with $N=2^r$. One could train a higher resolution quantizer, but test with various lower resolution systems without explicit retraining. By using the model downsampling framework, the system has access to multiple resolutions of HMM-based processing while only retaining one model in memory. This is especially important for distributed and/or resource-constrained applications.

In general, let a r -bit quantizer for channel m be comprised of centroids $\{c_m^{r,1}, c_m^{r,2}, \dots, c_m^{r,2^r}\}$. (Note that typically the original quantizer is allocated up to 8 bits ([PSS01],[PSP02][PSP03])). Similarly, the multi-resolution quantizations of clean speech are denoted by

$$q_m^r(A_m(n)) = c_m^{r,i}, \text{ where } i = \arg \min_j \|A_m(n) - c_m^{r,j}\|^2. \quad (6.29)$$

Using tree-structure quantization, centroids can be mapped according to

$$c_m^{8,i} \Rightarrow c_m^{r,j}, \text{ for } 1 \leq r < 8, \text{ where } j = \left\lfloor \frac{i}{2^{8-r}} \right\rfloor. \quad (6.30)$$

In Eq. 6.30, neighboring centroids are simply grouped in pairs as r decreases. In this manner, centroids can easily be regrouped as clusters without requiring expensive retraining of quantization codebook. Centroid boundaries, $z^{(i,i+1),r}$, are adapted similarly. Although the proposed framework is robust to the resolution chosen for τ , further derivations will set $\tau=8$.

6.4.2 Adaptation of Statistical Parameters

The signal model, now referred to as Ξ_m^r , has statistical parameters $(\mathbf{A}_m^r, \mathbf{B}_m^r, \boldsymbol{\pi}_m^r)$. The steady-state and transitional statistics can be approximated according to

$$\pi_i^{m,r} = \sum_{k=0}^{2^{8-r}-1} \pi_{(2^{8-r}i+k)}^{m,8}, \quad (6.31)$$

and

$$a_{ij}^{m,r} = \frac{1}{2^{8-r}} \left[\frac{\sum_{k=0}^{(2^{8-r}-1)} \pi_{(2^{8-r}i+k)}^{m,8} \sum_{l=0}^{(2^{8-r}-1)} a_{(2^{8-r}i+k, 2^{8-r}j+l)}^{m,8}}{\sum_{k=0}^{(2^{8-r}-1)} \pi_{(2^{8-r}i+k)}^{m,8}} \right]. \quad (6.32)$$

However, when quantizers are designed to minimize expected distortion, as is done in the current algorithm, the steady-state probabilities of quantizer centroids can be expected to be close to uniform. In this case, the formula in Equation 6.32 can be efficiently approximated as

$$a_{ij}^{m,r} = \frac{1}{2^{8-r}} \left[\sum_{k=0}^{(2^{8-r}-1)} \sum_{l=0}^{(2^{8-r}-1)} a_{(2^{8-r}i+k, 2^{8-r}j+l)}^{m,8} \right]. \quad (6.33)$$

During estimation of spectral components degraded by acoustic noise, the observation statistics $b_i^m(r_m(n))$ correspond to the probability that a clean spectral value belonging to centroid c_m^i is corrupted by additive noise and observed as $R_m(n)$. Hence, observation probability distributions of downsampled models are

$$b_i^{m,R}(R_m(n)) = p(R_m(n) | c_m^{R,i}) \quad (6.34)$$

6.4.3 Approximated HMM-Based Estimation

The HMM-based estimation techniques discussed in Section 6.3 can be applied to the downsampled statistical framework developed in Section 6.4. The estimate of component $R_m(n)$ using the approximation of the \mathbf{FB}_T algorithm with r bits of resolution becomes

$$\hat{A}_m^r(n) = \sum_{i=1}^{2^r} c_m^{r,i} \gamma_m^{r,i}(n), \quad (6.35)$$

where

$$\gamma_{T,m}^{r,i}(n) = \frac{\alpha_m^{r,i}(n) \beta_m^{r,i}(n)}{\sum_{j=1}^{2^r} \alpha_m^{r,j}(n) \beta_m^{r,j}(n)}. \quad (6.36)$$

The values $\alpha_m^{r,i}(n)$ and $\beta_m^{r,i}(n)$ can be determined recursively as

$$\alpha_m^{r,i}(n) = \left[\sum_{j=1}^{2^r} a_{ji}^{m,r} \alpha_m^{r,j}(n-1) \right] b_i^{m,r}(r_m(n)), \quad (6.37)$$

$$\beta_m^{r,i}(n) = \sum_{j=1}^{2^r} a_{ij}^{m,r} \beta_m^{r,j}(n+1) b_j^{m,r}(r_m(n+1)). \quad (6.38)$$

Similarly, $R_m(n)$ can be estimated using the \mathbf{FB}_F algorithm with r bits of resolution via Equation 6.35, where

$$\gamma_{F,m}^{r,i}(n) = \frac{\delta_m^{r,i}(n) \epsilon_m^{r,i}(n)}{\sum_{j=1}^{2^r} \delta_m^{r,j}(n) \epsilon_m^{r,j}(n)}. \quad (6.39)$$

In this case, the forward and backward variables are determined as

$$\delta_m^{r,i}(n) = \left[\sum_{j=1}^N P_{m:m-1}^{i:j} \delta_{m-1}^{r,j}(n) \right] b_i^{m,r}(r_m(n)), \quad (6.40)$$

$$\epsilon_m^{r,i}(n) = \sum_{j=1}^N P_{m:m+1}^{j:i} \epsilon_{m+1}^{r,j}(n) b_j^{m+1,r}(r_{m+1}(n)). \quad (6.41)$$

Finally, the \mathbf{FB}_{2D} algorithm estimates $R_m(n)$ via Equation 6.35, where

$$\gamma_{2D,m}^{r,i}(n) = \gamma_{T,m}^{r,i}(n) \gamma_{F,m}^{r,i}(n) = \frac{\alpha_m^{r,i}(n) \beta_m^{r,i}(n) \delta_m^{r,i}(n) \epsilon_m^{r,i}(n)}{\sum_{j=1}^{2^r} \alpha_m^{r,j}(n) \beta_m^{r,j}(n) \sum_{k=1}^{2^r} \delta_m^{r,k}(n) \epsilon_m^{r,k}(n)}. \quad (6.42)$$

Thus, HMM-based spectral reconstruction can be performed efficiently utilizing lower resolution statistical models, without expensive retraining or unnecessary memory requirements.

6.4.4 Performance and Complexity Analysis

The efficient HMM-based estimation techniques developed in Section 6.4 can intuitively be expected to degrade ASR performance as the resolution of underlying signal models is reduced. This section examines the extent to which this performance degradation occurs, and the possible reductions in computational complexity. Figure 6.6 illustrates the effect of model downsampling on word accuracy. Here, statistical masks are used (see Section 6.1 for details) in series with the \mathbf{FB}_T spectral reconstruction method, at various model resolutions. The system was tested on the Aurora database [Pea00], specifically the subset corrupted by vehicular noise, at various SNR levels. As can be concluded from Figure 6.6, the proposed model downsampling method leads to minimal performance degradation for resolutions as low as $r=3$.

The proposed downsampling methods can greatly reduce the required complexity

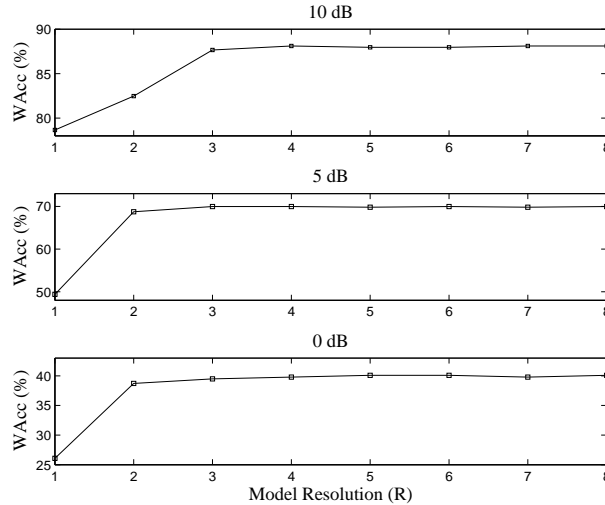


Figure 6.6: The effect of model downsampling on ASR performance: Statistical masks (see Chapter 6.1) were used in series with the \mathbf{FB}_T algorithm for various model resolutions.

of the proposed estimation system. Note that the order of complexity for the forward-backward decoding algorithm applied to a fully connected model is $O(2^{2N+1})$, where $N=2^r$ is the number of states comprising each underlying statistical model. The proposed estimation techniques were applied to a sample utterance, and the average number of operations per frame are provided in Table 8.1. Note that Table 8.1 excludes operations required by calculation of observation statistics, which are independent of model resolution. The sample utterance from the Aurora-2 database was 2.2 seconds in duration, and was degraded by vehicular noise at 15 dB SNR. As can be concluded, the proposed HMM downsampling method greatly reduces the induced computational load of spectral reconstruction. For example, if the resolution of underlying models is decreased from the standard 8 bits to 3 bits, the required complexity is reduced by a factor of over 800.

Based on the performance analysis of Figure 6.6 and the complexity analysis of Table 8.1, it can be concluded that the optimal tradeoff between recognition accuracy and induced computational load is obtained at a resolution of $r=3$. At this resolution, performance is affected negligibly relative to the standard 8-bit model system, at least for the

Table 6.2: Average operations per frame, as a function of model resolution, r . \mathbf{FB}_T refers to intra-channel HMM-based estimation, \mathbf{FB}_F refers to inter-channel HMM-based estimation, and \mathbf{FB}_{2D} refers to a combination of both. These results were obtained on a single utterance from the Aurora-2 database, with the window update set to 100 Hz, and using 26 Mel channels.

Operation	multiplications	additions	
r=8	\mathbf{FB}_T	4.342×10^5	4.299×10^5
	\mathbf{FB}_F	4.293×10^5	4.253×10^5
	\mathbf{FB}_{2D}	9.755×10^5	9.664×10^5
r=5	\mathbf{FB}_T	7,132	6,591
	\mathbf{FB}_F	7,005	6,591
	\mathbf{FB}_{2D}	1.594×10^4	1.479×10^4
r=3	\mathbf{FB}_T	520	383
	\mathbf{FB}_F	501	375
	\mathbf{FB}_{2D}	1,145	854

Aurora-2 database. However, in terms of operations required, the complexity is reduced by a factor of over 800, and in terms of processing time, the complexity is reduced by a factor of over 130.

The analysis provided in this section regarding model downsampling hints at the robustness of the recognizer to small changes in feature values. That is, the recognizer seems not to react to minor variability in features obtained from reconstructed spectrographic data at various resolutions. Instead, it appears that the general spectral shape of estimated data determines recognizer output.

6.5 Experimental Results

The overall noise robust recognition system featuring spectral reconstruction was tested on the Aurora-2 database [Pea00], as described in Chapter 6.1. The proposed framework was trained using a model with resolution of $r=8$, but was tested with a downsampled

model with $r=3$. This resolution was empirically shown in Section 6.4.4 to be the optimal tradeoff between performance and complexity for the given database. During testing, spectral reconstruction was only applied to those frames wherein active speech components were discovered during the mask estimation process, implicitly applying voice activity detection (VAD). Inactive frames were set to a spectral floor equal to $0.15R_m(n)$, similar to noise flooring described in [ZA03].

6.5.1 Recognition with Oracle Masks

Table 6.3 provides word-accuracy results for the proposed missing feature methods, when combined with oracle masks. The baseline, denoted as "none", performs recognition on unprocessed speech signals. Note that when oracle information is present, the derived noise robust ASR system suffers little degradation as noise levels increase.

It can be observed in Table 6.3 that exploiting inter-channel correlation provides better noise robust recognition than exploiting intra-channel correlation alone. On average, the \mathbf{FB}_F and \mathbf{FB}_{2D} algorithms provide the best results for the given database.

The use of oracle information in missing feature-based ASR provides an upper performance bound for a particular MF technique. If mask estimation methods used in this study are improved to better differentiate between reliable and unreliable spectrographic components, the overall recognition performance can be expected to approach the upper bound provided by Table 6.3.

Besides providing an upper performance bound, oracle masks allow decoupling of the mask estimation and data imputation components of missing feature systems. As in [CGJ01], the proposed spectral reconstruction technique is compared in the oracle mask case with data imputation based on spectral subtraction [Bol79]. Spectral subtraction imputation is defined as

Table 6.3: Word-accuracy results for HMM-based spectral reconstruction using oracle masks. "none" refers to unprocessed speech signals. \mathbf{FB}_T refers to estimation along the temporal axis, \mathbf{FB}_F refers to estimation along the frequency axis, and \mathbf{FB}_{2D} refers to a combination of both.

SNR (dB)	20	15	10	5	0	Ave
none	95.8	89.6	70.3	34.6	5.2	59.1
\mathbf{FB}_T	98.4	98.0	96.7	92.7	82.7	93.7
\mathbf{FB}_F	98.5	98.2	97.5	94.4	86.0	94.9
\mathbf{FB}_{2D}	98.6	98.2	97.2	94.0	86.3	94.9

$$\hat{A}_m(n) = \begin{cases} R_m(n), & \text{if } R_m(n) \text{ is deemed reliable} \\ \left(x_m^\tau(n) - \hat{D}_m^\tau(R_m(n))\right)^{1/\tau}, & \text{else} \end{cases}. \quad (6.43)$$

Here, τ is a user-defined parameter. Note that $\tau=2$ is equivalent to power spectral subtraction (PSS), and $\tau=1$ is equivalent to magnitude spectral subtraction (MSS). Figure 6.7 provides word-accuracies for the proposed \mathbf{FB}_F spectral reconstruction algorithm in the oracle mask case. Included for comparison are spectral subtraction-based imputation for $\tau=1$ and $\tau=2$. It can be observed from Figure 6.7 that the proposed method greatly outperforms the baseline techniques. Also, note that the proposed algorithm was initialized with the MSS estimate (Section 6.2). Thus, the increase in performance relative to SS with $\tau=1$ can be attributed directly to HMM-based reconstruction techniques.

6.5.2 Sensitivity Analysis for κ

Section 6.2 introduced a simple linear approximation of σ_ε , the standard deviation of noise estimation error, as a function of estimated noise magnitude, \hat{D}_m . The parameter σ_ε is important to the spectral reconstruction process since it is required during the calculation of observation statistics (Eq. 6.34). Section 6.2 used a small amount of noise data to set the constant κ to 0.6. Here, sensitivity analysis of κ is provided, and it is shown that the

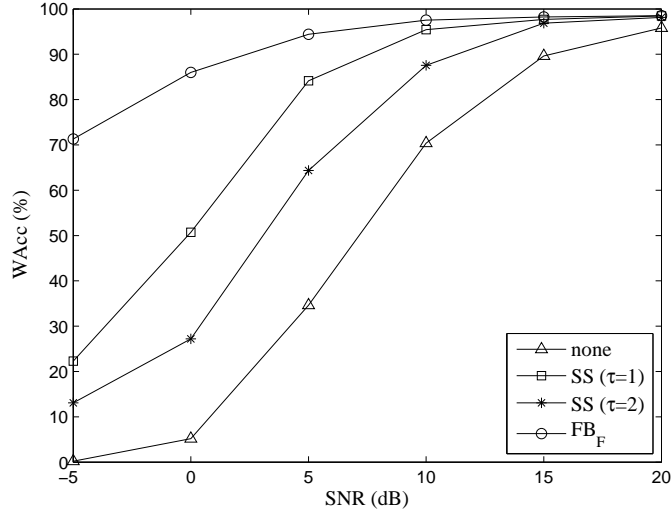


Figure 6.7: Word-accuracy results for \mathbf{FB}_F spectral reconstruction using oracle masks: SS refers to the spectral subtraction-based imputation technique defined in Eq. 6.43, and "none" refers to unprocessed signals. Results were averaged across all noise types in Set A of the Aurora-2 database.

exact value of this parameter is not important with respect to word-accuracy rate. Thus, extensive training is not necessary.

The proposed \mathbf{FB}_F missing feature-based ASR system was tested according to Section 1.3.3. The system was used in series with oracle masks. The original value of κ was altered and the resulting effects on word-accuracy rate was observed. Table 6.4 provides sensitivity analysis results for κ . $\Delta\kappa$ refers to the percent change in κ , the \overline{WAcc} refers to the word-accuracy averaged across all noise conditions, and $\Delta\overline{WAcc}$ refers to the change in average word-accuracy relative to that obtained for $\Delta\kappa=0$. It can be observed in Table 6.4 that changes in κ do not significantly effect the resulting recognition rate of the overall system, making extensive training of κ unnecessary.

6.5.3 Recognition with Statistical Masks

Table 6.5 provides word-accuracy results for the proposed missing feature methods, when combined with statistical masks from Section 6.1. Again, "none" refers to performing

Table 6.4: Sensitivity analysis for κ : $\Delta\kappa$ refers to the percent change in κ , \overline{WAcc} refers to the word-accuracy averaged across all noise conditions, and $\Delta\overline{WAcc}$ refers to the change in average word-accuracy relative to that obtained for $\Delta\kappa=0$.

$\Delta\kappa$	\overline{WAcc}	$\Delta\overline{WAcc}$
-30.0%	92.13	+1.24
-20.0%	91.90	+0.99
-10.0%	91.57	+0.63
$\pm 0.0\%$	91.00	± 0.00
+10.0%	90.37	-0.69
+20.0%	89.76	-1.36
+30.0%	89.07	-2.12

Table 6.5: Word-accuracy results for HMM-based spectral reconstruction using statistical masks from Chapter 6.1. "none" refers to unprocessed speech signals. \mathbf{FB}_T refers to estimation along the temporal axis, \mathbf{FB}_F refers to estimation along the frequency axis, and \mathbf{FB}_{2D} refers to a combination of both. Results were obtained at a resolution of $r=3$. Bold entries refer to the maximum results for each condition in the average case.

SNR (dB)	20	15	10	5	0	Ave.
none	95.8	89.6	70.4	34.6	5.2	59.1
\mathbf{FB}_T	97.7	95.3	88.5	74.0	49.7	81.0
\mathbf{FB}_F	97.7	95.3	89.3	75.6	50.1	81.6
\mathbf{FB}_{2D}	97.7	95.2	88.9	75.4	50.4	81.5

recognition on unprocessed speech signals. The resolution of the underlying statistical model was again set to $r=3$. As seen in Table 6.5, the proposed spectral reconstruction algorithms provide greatly improved recognition results for noisy conditions, relative to the baseline system.

Table 6.6: Word-accuracy results for HMM-based spectral reconstruction in the log-spectral domain, using statistical masks from Chapter 6.1

SNR (dB)	20	15	10	5	0	Ave.
none	95.80	89.64	70.38	34.61	5.17	59.1
FB_F (SPP)	98.08	96.41	91.96	80.59	53.92	84.2
FB_F (ORC)	98.37	97.75	96.32	92.64	81.79	93.4

6.6 Extension to Reconstruction in the Log-Spectral Domain

As mentioned previously, the proposed MF framework is robust to various frequency scales. Additionally, it is robust to various data types. Here, theory developed in the previous sections is applied to log Mel-filtered spectra, after liftering (see Chapter 6.1). Table 6.6 provides word-accuracy results for **FB_F** HMM-based spectral reconstruction in the log-spectral domain.

This chapter presents a novel HMM-based framework for estimation of unreliable spectrographic data. The proposed framework utilizes hidden Markov models to reconstruct corrupted spectral components based on reliable information, unreliable observations, and an underlying signal model, to improve noise robust speech recognition. Separate spectral reconstruction methods are derived which exploit intra-channel (across-time) correlation, inter-channel (across-frequency) correlation, or a combination of both.

CHAPTER 7

Utilizing Compressibility in Reconstructing Spectrographic Data

This chapter proposes a novel missing feature data estimation algorithm for noise robust ASR based on the notion of compressibility. Quantitative analysis is provided for the compressibility of spectrographic speech data, which motivates spectral reconstruction to be posed as an optimization problem minimizing the ℓ_1 -norm. A spectral reconstruction solution is presented, which is applied to front end MF-based ASR.

7.1 Signal Recovery from Incomplete Observations

Compressive sampling (CS) theory states that perfect reconstruction of signals can be achieved with far fewer observations than required by the traditional Nyquist sampling rate ([Don06],[CW08]). As discussed in [CW08], recovery of signals from an incomplete set of observations is made possible by the *sparsity* of the signal of interest, and by the *incoherence* of utilized sensing functions. A brief introduction to CS theory is presented, specifically regarding the notion of sparsity, providing motivation for the use of linear programming in the current problem of missing feature estimation.

Let $\mathbf{f} \in \mathbb{R}^N$ represent a signal of interest, and let the set $\varpi_k \in \mathbb{R}^N$, for $k = 1, \dots, M$, represent sensing functions used to obtain M observations in \mathbf{y} according to $\mathbf{y} = \Upsilon \mathbf{f}$, where Υ is comprised of row vectors ϖ_k . The design of sensing functions is an important aspect of many CS applications, such as imaging [Rom08], and involves the concept of

minimizing the coherence of bases (see [CW08] for details).

An underlying reason for the success of CS theory is that many signals can be described efficiently when expressed in terms of a proper basis. Let $\Omega \in \mathbb{R}^{N \times N}$ represent a suitable orthonormal basis for \mathbf{f} , such that $\mathbf{f} = \Omega^* \mathbf{v}$, where the $*$ operator represents the conjugate transpose. Here, \mathbf{v} is the sparse representation of \mathbf{f} , expanded in the basis Ω , also referred to as the representation basis. The compressible or sparse nature of \mathbf{v} states that it is comprised of only a few large magnitude terms, and implies that discarding the small terms will result in little or no distortion.

Define a signal as *S-sparse* if it contains at most S nonzero terms. Furthermore, let \mathbf{v}_S be the vector comprised of the S largest magnitude terms of \mathbf{v} , with the remaining terms set to zero. The recovered version of the original signal can be expressed as $\mathbf{f}_S = \Omega^* \mathbf{v}_S$. If the original signal is truly *S-sparse*, the reconstructed signal will be perfect. However, the original signal may be *compressible*, so that the magnitude of terms in \mathbf{v} decreases quickly, which will result in a reconstructed signal with little distortion.

The signal reconstruction \mathbf{f}_S can not generally be reproduced since it requires oracle information regarding the locations of large magnitude terms in \mathbf{v} . However, the discussion of sparsity motivates the use of the ℓ_1 -norm during signal recovery, since optimization problems which minimize the ℓ_1 -norm tend to solutions comprised of few nonzero terms [BV04]. In [Don06] and [CW08], the CS solution to the signal recovery problem is given by $\tilde{\mathbf{f}} = \Omega^* \tilde{\mathbf{v}}$, where $\tilde{\mathbf{v}}$ is determined by

$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to: } \mathbf{y} = \Upsilon \Omega^* \tilde{\mathbf{v}}. \quad (7.1)$$

Thus, the reconstructed signal $\tilde{\mathbf{f}}$, given an incomplete set of observations \mathbf{y} , is the function which minimizes the ℓ_1 -norm of the sparse representation $\tilde{\mathbf{v}}$. Although the cost function in Equation 7.1 is nonlinear, the problem statement can be rearranged as a linear program, and thus be solved quite efficiently [BV04].

7.2 The Compressibility of Spectrographic Speech Data

Compressive sensing has been successfully applied to both image compression and image denoising [CW08], [Rom08]. At the heart of these applications lies the fact that images typically have sparse representations when expanded on certain transform bases. This section explores the compressibility of spectrographic speech data. The discussion will motivate the use of linear programming in the proposed missing feature estimation algorithm.

Let \mathbf{x} denote the vector representation of $A_k(n) \forall n, k$, formed by lexicographic ordering. It is assumed that there exists an orthonormal basis Ω which reveals a concise representation of \mathbf{x} , namely \mathbf{v} . Additionally, assuming oracle information regarding the location of large magnitude terms within \mathbf{v} , the S -sparse vector, \mathbf{v}_S , is extracted, from which the approximation of the original speech signal, \mathbf{x}_S , can be recovered. Let β be the portion of terms within \mathbf{v} retained, and be defined as $\beta = \frac{N_s}{N}$, where N_s is the number of nonzero terms retained, and N is the total number of elements. The quality of the recovered signal, as a function of β , can be analyzed to assess the compressibility of the original data.

In image processing, the mean-square error (MSE) distortion provides a reliable metric for measuring the degradation of a compressed image [Jai89]. However, in speech processing, such a distance metric applied to spectrographic speech data does not directly reflect the quality of the underlying speech data, and instead the performance of the overall speech processing system must be analyzed. In automatic speech recognition (ASR), one can study the effect of induced sparsity on the resulting recognition accuracy rates.

Figure 7.1 shows an example of the sparsity of Mel-filtered spectrographic speech data. Analysis was performed on the word "three" extracted from the Aurora-2 database. The top panel shows the sparse representation of the input Mel-filtered spectrographic data in vector form, utilizing the discrete Haar transform (DHT) [Jai89]. The DHT was

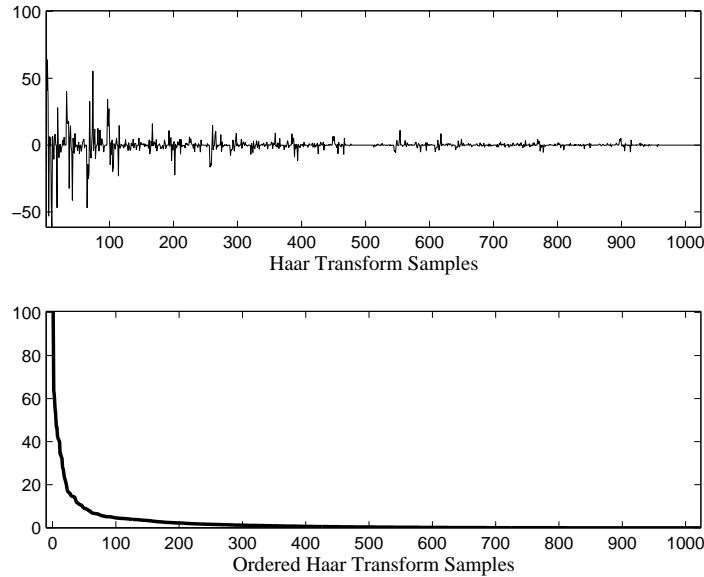


Figure 7.1: The compressibility of spectrographic speech data: Analysis was performed on the clean word "three" extracted from the Aurora-2 database [Pea00]. The top panel shows the sparse representation of the input spectrographic data in vector form, utilizing the discrete Haar transform (DHT). The bottom panel shows the absolute value of the sparse representation, sorted by magnitude.

chosen due to its common usage in compressive sensing. Other transforms such as the Discrete Cosine Transform (DCT) and the Karhunen-Loeve Transform (KLT) were tested, but with less success. The bottom panel shows the absolute value of the sparse representation, sorted by magnitude. As can be concluded from the rapidly decreasing values in the bottom panel, the input spectrographic data are highly compressible.

Figure 7.2 provides quantitative analysis of the compressibility of Mel-filtered spectrographic speech data, presenting the time-average MSE, along with word-accuracies, as a function of induced sparsity. The representation basis used was again the discrete Haar kernel. Compressibility analysis was performed on clean speech from the Aurora-2 database [Pea00]. It can be concluded from Figure 7.2 that approximately 85% ($\beta=0.85$) of terms in the sparse domain can be zeroed without significantly affecting recognition results.

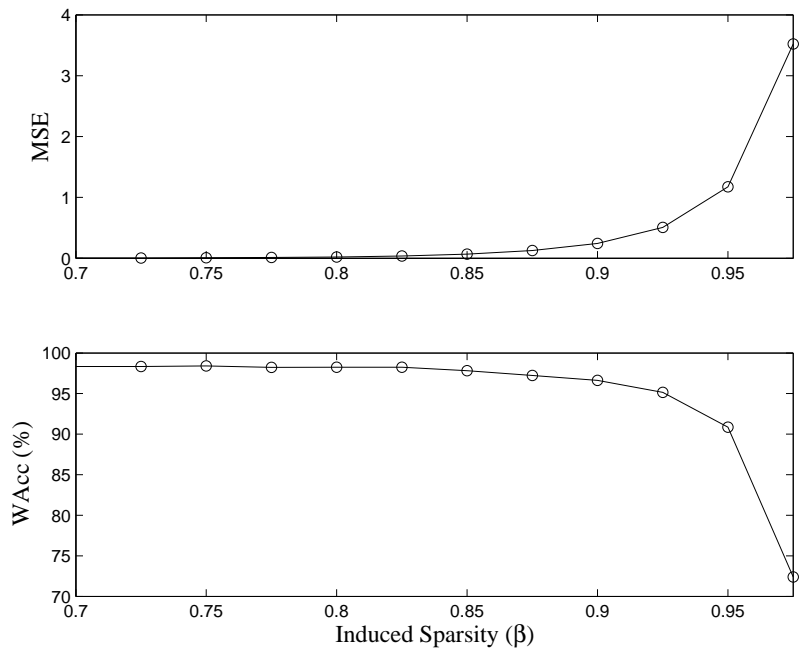


Figure 7.2: Quantitative analysis of the compressibility of spectrographic speech data: The top panel illustrates the MSE distortion resulting from induced sparsity in \mathbf{x}_S , utilizing the discrete Haar transform (DHT). The bottom panel provides word-accuracies corresponding to the recovered Mel-filtered spectra used in the top panel.

7.3 Reconstruction of Missing Features for Noise Robust Speech Processing

In real world applications, speech signals will generally suffer degradation due to acoustic noise, resulting in decreased performance for recognition. This section presents a missing feature estimation method for noise suppression of spectrographic speech data, which in turn is applicable to automatic speech recognition.

7.3.1 The Proposed Missing Feature Estimation Algorithm

Let $\Omega \in \mathbb{R}^{N \times N}$ be the representation basis revealing a compressible representation of \mathbf{x} , namely \mathbf{v} . In MF-based noise robust ASR systems, spectral reconstruction algorithms must be preceded by mask estimation. Two types of binary masks are utilized, oracle masks and ones based on speech presence probability (SPP), which classify each term in \mathbf{x} as reliable, corresponding to strong speech signal presence relative to noise, or unreliable, corresponding to a high level of corruption due to noise.

A spectral reliability mask can be expressed analytically via the selection matrix $\mathbf{H}_R \in \mathbb{R}^{M \times N}$, corresponding to the M reliable components of \mathbf{x} . \mathbf{H}_R is defined as follows

$$\mathbf{H}_R(i, j) = \begin{cases} 1, & \text{if } \mathbf{x}(j) \text{ is the } i^{th} \text{ reliable term in } \mathbf{x} \\ 0, & \text{otherwise} \end{cases}. \quad (7.2)$$

The goal of signal reconstruction can be restated as estimating the components of the sparse representation \mathbf{v} , given the incomplete reliable observations $(\mathbf{H}_R \mathbf{x})$. Motivated by the discussions on CS from Section 7.1, and on the compressibility of speech from Section 7.2, the missing feature estimation task can be posed as an optimization problem minimizing the ℓ_1 -norm

$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to:} \quad \mathbf{H}_R \mathbf{x} = \mathbf{H}_R \Omega^* \tilde{\mathbf{v}} \quad (7.3)$$

Utilizing Equation 7.3, the sparse representation of the estimated underlying speech spectrum can be determined. The reconstructed clean speech spectrum can be found as $\hat{\mathbf{x}} = \Omega^* \tilde{\mathbf{v}}$.

The spectral reconstruction solution expressed in Equation 7.3 does not take into account any information specific to spectrographic speech data. Basic properties of spectral signals can be integrated as constraints in the optimization problem of Equation 7.3 to better estimate the underlying clean speech spectrum. First, spectral coefficients are inherently nonnegative. Also, following the additive model from Equation 1.3, it can be concluded that clean speech components must be less than or equal to observed spectral components. Thus, the optimization from Equation 7.3 becomes

$$\begin{aligned} \min_{\tilde{\mathbf{v}} \in \mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to:} \quad & \mathbf{H}_R \mathbf{x} = \mathbf{H}_R \Omega^* \tilde{\mathbf{v}} & (7.4) \\ & \Omega^* \tilde{\mathbf{v}} \geq 0 \\ & \Omega^* \tilde{\mathbf{v}} \leq \mathbf{x} \end{aligned}$$

The additional constraints of Equation 7.4 provide boundaries for the solution $\tilde{\mathbf{v}}$, specific to Mel-filtered spectral data, which may not exist for other types of spectrographic data. These constraints were found to result in more accurate solutions than cases with no such constraints.

As stated previously, the optimization in Equation 7.4 can be carried out efficiently

by posing the problem as a linear program (LP) [BV04]. In this study, the primal-dual LP method was used during optimization. Also, the algorithm is able to run in real-time by processing one spectral frame (25 ms with a 10 ms overlap) at a time. Thus, the reconstructed log-spectral signal in the sparse domain is that which minimizes the ℓ_1 -norm, fixed by observed reliable log-spectral components, and constrained by upper and lower boundaries (ii) and (iii).

It is interesting to note that the subspace approach to noise suppression [HWh07] also exploits the sparsity of speech for noise-robust processing. Such techniques explicitly construct the sparse design of speech signals given predetermined threshold(s). In the proposed algorithm, however, the sparse design is determined implicitly during the minimization of the ℓ_1 -norm of the sparse representation $\tilde{\mathbf{v}}$.

7.3.2 Comparisons with Compressive Sensing

As can be interpreted from Equations 7.1 and 7.3, great similarity exists between compressive sensing and the proposed missing feature estimation algorithm. Both techniques aim to reconstruct signals from an incomplete set of observations. Additionally, both techniques rely on the notion of compressibility, and specifically each method minimizes the ℓ_1 -norm of the given signal in a sparse domain. However, the concept of sensing is quite different in each case.

In CS applications, such as imaging [Rom08], signals are sampled at low rates by utilizing sensing functions. Sensing functions are designed to comprise an orthonormal basis, Υ , which minimizes the coherence measure with the representation basis, Ω . (See [CW08] for a detailed discussion.)

In the proposed missing feature estimation algorithm, observations are not sampled in the same sense as in CS applications, but their origins are instead decided by the reliability of terms in the mask estimation domain. The sensing matrix, which for the proposed

algorithm can be written as $\Upsilon = \mathbf{H}_R$, is defined entirely by the corruptive effect of noise on the input speech signal. Thus, sensing functions cannot be actively designed, and the notion of coherence does not play a role.

7.4 Extension to Reconstruction in the Log-Spectral Domain

Motivated by observations from Section 7.3, the spectral reconstruction solution is adapted to the log Mel-spectral domain. \mathbf{x}_L , \mathbf{y}_L , and \mathbf{v}_L are defined as the log-domain versions of \mathbf{x} , \mathbf{y} , and \mathbf{v} , respectively

$$\mathbf{y}_L = \mathbf{H}_R \mathbf{x}_L, \quad (7.5)$$

and

$$\mathbf{v}_L = \Omega \mathbf{x}_L. \quad (7.6)$$

Note that constraint (ii) from Eq. 7.4 is valid for log-spectral data due to the monotonic nature of the logarithm. Constraint (iii), on the other hand, does not hold since log-spectral values become unbounded with respect to the lower limit. By applying log-spectral flooring from Chapter 6.1, constraint (iii) can be stated alternatively, and the data inference solution in the log-spectral domain can be expressed as

$$\begin{aligned} \min_{\tilde{\mathbf{v}}_L \in \mathbb{R}^N} \|\tilde{\mathbf{v}}_L\|_1 \quad \text{s.t.} \quad & (i) \mathbf{H}_R \mathbf{x}_L = \mathbf{H}_R \Omega^* \mathbf{v}_L \\ & (ii) \Omega^* \mathbf{v}_L \leq \mathbf{x}_L \\ & (iii) \Omega^* \mathbf{v}_L \geq \alpha_{fl} \end{aligned} \quad (7.7)$$

Table 7.1: Word-accuracies for the proposed missing feature reconstruction technique, using oracle masks (ORC) and statistical masks (SPP) from Chapter 6.1

SNR (dB)	20	15	10	5	0	Ave.
baseline	95.8	89.6	70.4	34.6	5.2	59.1
ORC	98.7	98.4	97.0	92.2	81.1	93.5
SPP	97.7	95.8	88.9	72.3	42.4	79.4

7.5 Experimental Results

The proposed spectral reconstruction methods are applied to the Aurora-2 database, using the experimental procedure described in Chapter 6.1. Table 7.1 provides word-accuracy rates for the proposed algorithm when combined with oracle reliability masks (ORC) and statistical masks (SPP) from Chapter 6.1. The proposed spectral reconstruction algorithm is shown to provide a high upper performance bound when combined with oracle reliability masks. Additionally, when combined with SPP-based masks, the proposed algorithm provides significant improvements in word-accuracy rates, relative to the baseline system which uses a standard MFCC front end.

Table 7.2 provides results for the CS-based reconstruction method, when applied in the log-spectral domain. It can be observed that application of the proposed MF technique on log Mel-filtered spectra provides a significant improvement in word-accuracy, relative to the Mel-filtered domain. Furthermore, the proposed method outperforms the ETSI advanced front end (AFE) [Doc07].

This chapter presents a novel algorithm for the reconstruction of unreliable spectral speech components based on the notion of compressibility. We provide quantitative analysis on the compressibility of spectrographic speech data, which motivates the use of minimization of the ℓ_1 -norm in the proposed missing feature estimation technique. The proposed spectral reconstruction method is shown to provide significant improvements in word-accuracy rates relative to the MFCC baseline system, when applied to the Aurora-2

Table 7.2: Word-Accuracies for the the proposed missing feature reconstruction technique applied in the log-spectral domain, using oracle masks (ORC) and statistical masks (SPP) from Chapter 6.1

SNR (dB)	20	15	10	5	0	Ave.
none	95.8	89.6	70.4	34.6	5.2	59.1
LS-FLR	97.6	95.2	88.2	69.6	34.2	77.0
SPP	98.3	97.2	94.1	84.5	60.3	86.9
ORC	98.7	98.0	96.4	91.6	78.7	92.7
ETSI AFE [Doc07]	98.2	97.0	92.2	79.1	51.1	83.5

database.

CHAPTER 8

Extension of the Missing Feature Approach to Packet Loss Concealment

Missing feature estimation techniques proposed in Chapters 6 and 7 are robust to various types of data. In this chapter, theory previously developed in Chapter 6 is extended to estimation of missing speech features during packet loss concealment.

In digital speech communication systems, transmitted data may become lost or corrupted due to channel degradation. Specifically, transmission of speech over wireless communication systems relies on error detecting codes to determine the reliability of received frames [FV01]. Packet-based systems, on the other hand, are subject to arrival jitter, which may induce unacceptable delay for real-time speech applications [RMA06]. In either scenario, packet loss concealment (PLC) is applied at the receiver to reconstruct perceptually valid speech frames.

To reduce the impact of lost frames, some studies have applied waveform substitution or extrapolation techniques in the time domain [CL07]. Other studies have performed PLC for parametric coders in the parameter domain ([RMA06],[Mar01]). This section presents novel efficient HMM-based techniques for estimating missing speech features, with applications to packet loss concealment. Speech parameters are assumed to be observations of hidden Markov processes, and generalized MMSE estimates of missing features are derived. In this way, the natural progression of speech features in time is capturing by exploiting *a priori* steady-state and transitional statistics. Furthermore, methods are offered by which to increase the efficiency of the proposed framework. Specifically,

underlying Markov models are downsampled, similar to [BA10a]. Additionally, symmetry of transitional probability matrices is enforced, allowing efficient computation.

8.1 HMM-Based Estimation

8.1.1 Interpreting Speech Parameters as Markov Processes

Parametric coders transmit feature vectors comprised of speech parameters which are used to synthesize speech at the receiver. Transmitted packets typically include spectral shape (such as line spectral frequencies), gain, and pitch information [Kon06]. This section derives estimation techniques for a general missing feature, x_n , where n denotes time index. Note that as opposed to HMM-based inference discussed in Section 6, the scenario of lost or delayed packets does not provide access to unreliable observations. Therefore, HMM observation statistics must be assumed uniform. Although this can be expected to degrade estimation performance, it does allow for efficient implementation, as well as concise vector-form notation during derivation of solutions.

Packet loss concealment (PLC) for parameter-based coders in the feature domain can be generalized as estimating x_{n+k} , conditioned on reliable features x_n and x_{n+N} , and given that $[x_{n+1}, \dots, x_{n+N-1}]$ are missing. For some PLC applications, future features (x_{n+N}) may not be available, since this may induce unacceptable delays [CL07]. In this section, however, missing feature estimation techniques are derived in the general case, where both past and future samples are available. The general case solution can then be reduced to the specific solution based solely on past observations.

Speech parameters are interpreted as observations of Markov processes, as in [RMA06]. In such a framework, parameter x_n is the observation of a fully connected hidden Markov model (HMM) with K states. State s_i is assumed to exhibit a continuous observation probability distribution function (pdf) that follows a normal distribution with mean μ_i .

Given the Markov property, the transitional probability between states at time index m can be expressed as

$$a_{ij} = P(s_m = i | s_{m-1} = j). \quad (8.1)$$

The matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ is defined such that $\mathbf{A}(i, j) = a_{ij}$.

Furthermore, the matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$ is defined such that

$$\mathbf{B}(i, j) = b_{ij}, \text{ where: } b_{ij} = P(s_m = i | s_{m+1} = j). \quad (8.2)$$

Parameter values are assumed to be outputs of a Markov process. HMMs with continuous observation pdfs are specified for codec parameters with a continuous range. For such features, it may be most suitable to apply the minimum mean-square error (MMSE) estimate, defined as [Rab89]

$$\hat{x}_{n+k} = \sum_{i=1}^K \mu_i P(s_{n+k} = i | x_n, x_{n+N}). \quad (8.3)$$

8.1.2 Deriving State-Specific Probabilities

This section provides derivations for state-specific conditional probabilities

$P(s_{n+k} = i | x_n, x_{n+N})$, which are required by the MMSE solution. Using a Bayesian approach, the conditional probabilities in Eq. 8.3 can be expressed as

$$\begin{aligned} P(s_{n+k} = i | x_n, x_{n+N}) &= \frac{P(s_{n+k} = i, x_n, x_{n+N})}{P(x_n, x_{n+N})} \\ &= \frac{P(s_{n+k} = i, x_n) P(x_{n+N} | s_{n+k} = i, x_n)}{P(x_n, x_{n+N})} \end{aligned} \quad (8.4)$$

Using the Markov property previously assumed, the left-hand probability in the numerator of Eq. 8.4 can be approximated as

$$P(s_{n+k} = i, x_n) = \begin{cases} \sum_{j=1}^T P(s_{n+k} = i | s_{n+k-1} = j) P(s_{n+k-1} = j, x_n), & \text{for } k > 0, \\ P(s_n = i | x_n) P(x_n), & \text{for } k = 0 \end{cases}$$

The hidden state at time n is inferred by minimizing the distortion between the observed reliable feature and the underlying model

$$P(s_n = i | x_n) = \delta_{i, q_n}, \text{ where } q_n = \arg \min_j \|x_n - \mu_j\|^2. \quad (8.5)$$

Thus, Eq. the left-hand probability in the numerator of Eq. 8.4 can be simplified as

$$P(s_{n+k} = i, x_n) = P(x_n) \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)}, \quad (8.6)$$

where the vector $\mathbf{e}_j \in \mathbb{R}^K$ is comprised of zeros, except for a one at the j^{th} element. As in the previous equation, the notation $[\mathbf{m}]_{(j)}$ will be used to refer to the j^{th} element of vector \mathbf{m} . The right-hand probability in the numerator of Eq. 8.4 can be approximated as

$$P(x_{n+N} | s_{n+k} = i, x_n) \approx P(x_{n+N} | s_{n+k} = i) = \frac{P(s_{n+k} = i, x_{n+N})}{P(s_{n+k} = i)} \quad (8.7)$$

Note that the denominator of Eq. 8.7 is the steady-state probability of state i , denoted by $\pi(i)$. Steady-state statistics can be determined as the elements of the eigenvector of \mathbf{A} or \mathbf{B} corresponding to the unit eigenvalue

$$\mathbf{A}\boldsymbol{\pi} = \boldsymbol{\pi}, \text{ and } \mathbf{B}\boldsymbol{\pi} = \boldsymbol{\pi}. \quad (8.8)$$

Furthermore, the numerator of Eq. 8.7 can be simplified by assuming the Markov property

$$P(s_{n+k} = i, x_{n+N}) = \begin{cases} \sum_{j=1}^K P(s_{n+k} = i | s_{n+k+1} = j) P(s_{n+k+1} = j, x_{n+N}), & \text{for } k < N \\ \delta_{i, q_{n+N}} P(x_{n+N}), & \text{for } k = N \end{cases} \quad (8.9)$$

Thus Eq. 8.7 can simplified as

$$P(x_{n+N} | s_{n+k} = i) = \frac{P(x_{n+N})}{\boldsymbol{\pi}(i)} [\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}}]_{(i)}. \quad (8.10)$$

By substituting Eqs. 8.6 and 8.10 into Eq. 8.4, the underlying state for time index $n + k$ is inferred via

$$P(s_{n+k} = i | x_n, x_{n+N}) = \frac{P(x_n) P(x_{n+N})}{P(x_n, x_{n+N})} \frac{1}{\boldsymbol{\pi}(i)} [\mathbf{A}^{(k)} \mathbf{e}_{q_n}]_{(i)} [\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}}]_{(i)}. \quad (8.11)$$

The first term, referred to as κ , is independent of state i , and is simply used to normalize the probability distribution

$$\kappa = \left(\sum_{i=1}^K \frac{1}{\boldsymbol{\pi}(i)} [\mathbf{A}^{(k)} \mathbf{e}_{q_n}]_{(i)} [\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}}]_{(i)} \right)^{-1} \quad (8.12)$$

The value κ need not be explicitly determined during estimation of x_{n+k} . Instead, κ cancels from the solution by assuring that

$$\sum_{i=1}^K P(s_{n+k} = i | x_n, x_{n+N}) = 1. \quad (8.13)$$

If future reliable features are not available, the underlying state of a missing feature is inferred by marginalizing the observation x_{n+N}

$$\begin{aligned} P(s_{n+k} = i | x_n) &= \int_{x_{n+N}} P(s_{n+k} = i | x_n, x_{n+N}) \partial x_{n+N} \\ &= \int_{-\infty}^{\infty} \frac{\kappa}{\boldsymbol{\pi}(i)} \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)} \left[\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}} \right]_{(i)} \partial x_{n+N} \\ &= \frac{\hat{\kappa}}{\boldsymbol{\pi}(i)} \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)} \left[\mathbf{B}^{(N-k)} \boldsymbol{\pi} \right]_{(i)} \\ &= \hat{\kappa} \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)}, \end{aligned} \quad (8.14)$$

$$\text{where: } \hat{\kappa} = \left(\sum_{i=1}^K \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)} \right)^{-1} \quad (8.15)$$

Depending on the availability of future observations, \hat{x}_{n+k} is determined by substituting either Eq. 8.11 or 8.14 into Eq. 8.3.

Note that [RMA06] offers a similar derivation for estimation of missing features. However, several important differences exist between the two approaches. In [RMA06], determining state-specific probabilities of the reliable "boundary" feature requires HMM-based decoding of a series of preceding features. In the proposed method, such probabilities are determined simply by Eq. 8.5. Furthermore, in the proposed work, Eq. 8.11 includes the steady-state probability in the denominator, which is missing from the corresponding equation in [RMA06]. Finally, this section provides the solution for the scenario in which future observations are not available via marginalization of a generalized solution conditioned on past and future observations.

8.2 Reducing the Complexity of HMM-Based Estimation

8.2.1 Markov Model Downsampling

From Eq. 8.11, it can be observed that the size of underlying signal models, K , has a large effect on the resulting complexity of the proposed algorithm. As was previously explored in Section 6, the configuration of underlying Markov models can be downsampled to increase the computational efficiency of HMM-based estimation. Underlying models can either be trained at multiple resolutions, or lower resolution statistics can be extracted from a higher order model. Details are provided in Section 6.

8.2.2 Enforcing Transition Matrix Symmetry

From Section 8.1, the HMM-based estimation in the general missing feature framework is given by

$$P(s_{n+k} = i | x_n, x_{n+N}) = \frac{\kappa}{\pi(i)} \left[\mathbf{A}^{(k)} \mathbf{e}_{q_n} \right]_{(i)} \left[\mathbf{B}^{(N-k)} \mathbf{e}_{q_{n+N}} \right]_{(i)}. \quad (8.16)$$

Note that Eq. 8.16 requires numerous self-multiplications of transitional matrices \mathbf{A} and \mathbf{B} . For large underlying Markov models (i.e. large K), this may prove computationally expensive. However, certain statistical patterns of transitional matrices can be exploited to reduce the complexity of Eq. 8.16.

Suppose \mathbf{A} is symmetric. Its singular value decomposition (SVD) then reveals equivalent input and output bases: $\mathbf{A} = \Phi_a \Lambda_a \Phi_a^*$, where Λ_a is diagonal and Φ_a is orthonormal. The self-multiplication of \mathbf{A} can then be expressed as

$$\mathbf{A}^k = (\Phi_a \Lambda_a \Phi_a^*)^k = \Phi_a \Lambda_a^k \Phi_a^*, \quad (8.17)$$

reducing the required computation due to the diagonal nature of Λ_a . If \mathbf{A} and \mathbf{B} are

symmetric matrices, Eq. 8.16 becomes

$$P(s_{n+k} = i | x_n, x_{n+N}) = \frac{\kappa}{\pi(i)} [\Phi_a \Lambda_a^k \Phi_a^* \mathbf{e}_{q_n}]_{(i)} [\Phi_b \Lambda_b^{N-k} \Phi_b^* \mathbf{e}_{q_{n+N}}]_{(i)}. \quad (8.18)$$

Transitional matrices of LSF parameters show strong symmetric patterns. However, because such matrices are data-generated, they are not perfectly symmetric, and a perturbation matrix Δ can be added to \mathbf{A} so that

$$\bar{\mathbf{A}} = \mathbf{A} + \Delta = (\mathbf{A} + \Delta)^T = \bar{\mathbf{A}}^T. \quad (8.19)$$

Because \mathbf{A} is a transition probability matrix, its columns each sum to unity, i.e. $\mathbf{1}^T \mathbf{A} = \mathbf{1}^T$, where $\mathbf{1}$ is an appropriately-sized unity vector. Additionally, Δ should be as small as possible to minimize its statistical effect on \mathbf{A} . This leads to

$$\begin{aligned} \dot{\Delta} = \arg \min_{\Delta} \|\Delta\|_{FW} \text{ s.t. } (i) \mathbf{1}^T \Delta = \mathbf{0}^T \\ (ii) \mathbf{A} + \Delta = (\mathbf{A} + \Delta)^T. \end{aligned} \quad (8.20)$$

Here, $_{FW}$ indicates a weighted Frobenius norm where each entry in Δ can be weighted differently to avoid negative entries in $\bar{\mathbf{A}}$. Δ_{ij} is weighted by $\frac{1}{\mathbf{A}_{ij} + \mathbf{A}_{ji}}$, thereby restricting large changes in near-zero entries of \mathbf{A} . Eqn. (8.20) can be posed as

$$\dot{\mathbf{d}} = \arg \min_{\mathbf{d}} \|\mathbf{d}\|_{FW} \text{ s.t. } (i) \mathbf{A}_{eq} \mathbf{d} = \mathbf{b}_{eq}, \quad (8.21)$$

where $\mathbf{d} = \text{vec}(\Delta)$, and where \mathbf{A}_{eq} and \mathbf{b}_{eq} represent a set of linear equality constraints

on \mathbf{d} imposed by (i) and (ii) from Eq. 8.20. The first constraint in Eqn. 8.20 results in a set of n equality constraints, and the second yields another $\frac{n^2-n}{2}$ constraints. Thus, $\mathbf{A}_{eq} \in \mathbb{R}^{\frac{n^2+n}{2} \times n^2}$ represents an underdetermined linear system, and $\mathbf{A}_{eq}\mathbf{d} = \mathbf{b}_{eq}$ has an infinite set of solutions. The weighted least-norm solution to this underdetermined set of equations is well-known

$$\dot{\mathbf{d}} = \mathbf{W}^{-1} \mathbf{A}_{eq}^T (\mathbf{A}_{eq} \mathbf{W}^{-1} \mathbf{A}_{eq}^T)^{-1} \mathbf{b}_{eq}, \quad (8.22)$$

where \mathbf{W} is a weighting matrix and $\dot{\mathbf{\Delta}}$ is obtained by reshaping $\dot{\mathbf{d}}$. A corresponding perturbation matrix for \mathbf{B} can be found similarly. It is important to note that the singular value decompositions and optimization techniques proposed in this section are performed offline, and required complexity of these operations is therefore not an issue.

8.2.3 Complexity Analysis

Sections 8.2.1 and 8.2.2 present methods by which to reduce the complexity of the original HMM-based estimation method of Eq. 8.11, without noticeable degradation in performance. This section provides quantitative complexity analysis, and show the efficient method of Eq. 8.18 to result in significant reductions in required computation for error burst lengths of ≥ 2 .

The computational complexity associated with matrix multiplication of full matrices of size $M \times M$ is known to be $O(M^3)$. If either of the matrices is known to be diagonal, this sparse structure can be exploited, and the complexity reduces to $O(M^2)$. If both matrices are diagonal, the complexity reduces further to $O(M)$. Using this, the number of required multiplications for the standard method of Eq. 8.11 and reduced complexity method of Eq. 8.18 are plotted in Figure 8.1 as a function of the duration of error burst. Note that r refers to the resolution of the downsampled HMM. It is clear from Figure 8.1 that model downsampling and enforcing matrix symmetry offers significantly reduced

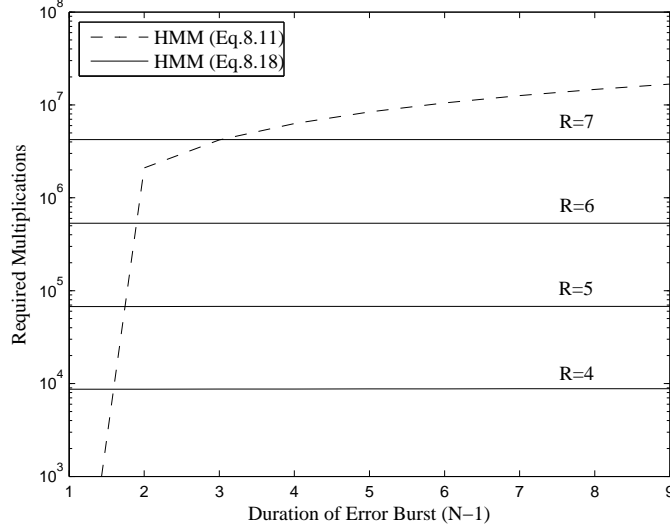


Figure 8.1: Induced complexity of HMM-based estimation of missing features for original method (Eq. 12) and reduced complexity method (Eq. 19)

complexity for error durations of ≥ 2 , making the algorithm ideal for bursty channels. In fact, PLC schemes can be implemented wherein Eq. 8.18 is applied only for missing features occurring at least 2 frames into an error burst.

8.3 Experimental Results

The proposed estimation techniques were applied to speech features generally utilized by parametric speech coders, namely line spectral frequencies (LSFs). A bursty channel is simulated, for which a two-state model is used, wherein state 0 incurred no loss, and state 1 incurred a dropped packet with probability 1. The probability of self-transition within state 1 was twice that of the probability of transition from 0 to 1.

The proposed HMM-based framework was applied to 100 randomly selected utterances from the TIMIT database, separate from the training set, and using the previously mentioned channel model. Figure 8.2 provides an illustrative example of the reconstruction of missing values for the 1st LSF. "REP" refers to the baseline scheme wherein

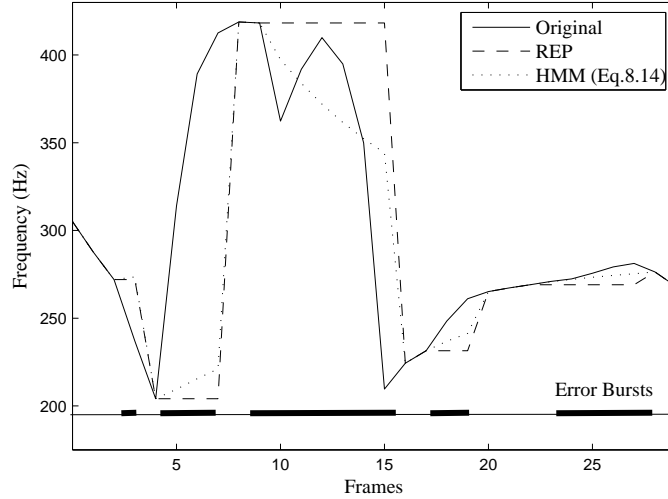


Figure 8.2: Reconstructed trajectories for the 1st LSF in the presence of error bursts. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to HMM-based estimation with $r=7$. Error bursts are denoted by horizontal bars.

reliable features are repeated (as in [Mar01]), whereas "HMM" refers to the proposed HMM-based framework. It can be observed that proposed HMM-based estimation generally provides more accurate reconstructions, as well as smoother transitions, relative to the baseline. The quantitative quality of reconstructed LSF feature trajectories was assessed by peak-SNR (PSNR) and weighted LSF distortion (WSD) [PA93]. Figure 8.3 and Table 8.1 provide results for $r=7$. Recall that the difference between Eq. 8.14 and 8.11 is that the latter requires future observations, whereas the former does not. It can be observed that both HMM-based methods provide significant improvements relative to the baseline, in terms of PSNR and WSD. Note that WSD incorporates a model of the human auditory system, and can therefore be expected to exhibit correlation with perceptual quality.

Figure 8.4 provides results for estimation of missing LSF vectors using reduced complexity estimation developed in Section 8.2. Specifically, the figure illustrates the effect of HMM downsampling on WSD, for a 5.0% error rate. It can be observed that performance of the proposed estimation technique converges for $r \approx 4$, corresponding to 16 states. Note that this corresponds to a model size that is significantly smaller than those used

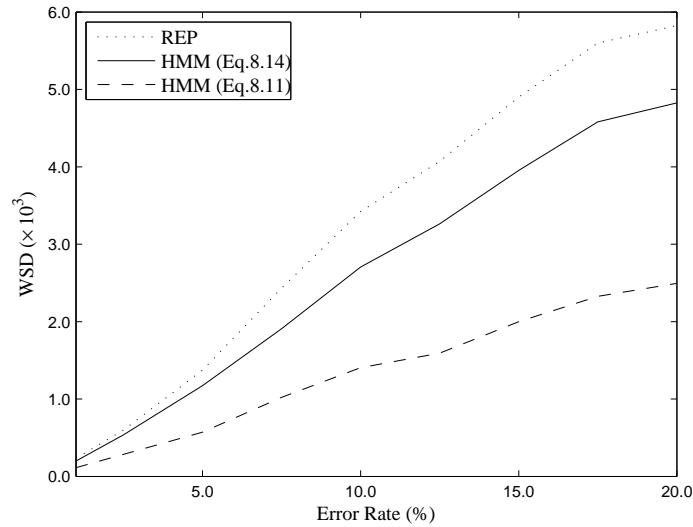


Figure 8.3: Weighted LSF distortion [PA93] for estimation of missing LSF features as a function of error rate. "REP" refers to the baseline repetition scheme, whereas "HMM" refers to the proposed HMM-based framework with $r=7$.

Table 8.1: Improvements in Peak-SNR (dB), relative to feature repetition, for Estimation of Missing LSF Features as a Function of Error Rate, for $r=7$. Results are Averaged Across 10 Individual LSFs.

Error Rate (%)	5	10	15	20
HMM (Eq. 8.14)	0.75	1.06	0.96	0.84
HMM (Eq. 8.11)	3.62	3.62	3.64	3.46

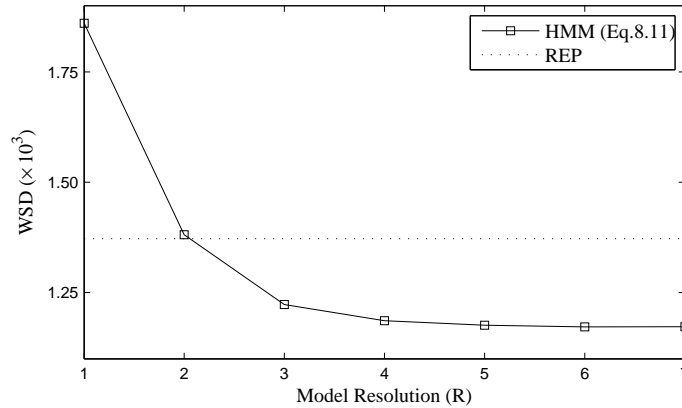


Figure 8.4: The effect of model downsampling on WSD for HMM-based estimation with Eq. 8.11 with a 5% error rate

previously in [RMA06]. Furthermore, it was observed that enforcing transition matrix symmetry had a negligible effect on estimation performance in terms of WSD.

This chapter presents efficient HMM-based estimation techniques for missing speech features, with applications to parametric coding. By assuming features to be observations of hidden Markov processes, the minimum mean-square error solutions for estimating unreliable features is derived. Computationally efficient approximations to the derived solutions are explored by downsampling underlying Markov models, and by enforcing symmetry in transitional probability matrices. When applied to features generally utilized by parametric coding, the proposed estimation methods outperform baseline repetition scheme in terms of both PSNR and WSD.

CHAPTER 9

Summary and Future Work

In real world speech processing systems, speech signals are often corrupted by background acoustic noise or reverberation. Additionally, for systems which involve transmission of speech data over error-prone communication channels, signals may suffer from packet loss. This dissertation addresses two general frameworks for which compensation of corruptive acoustic noise and channel errors can benefit performance, namely remote speech communication and automatic speech recognition.

9.1 Short-Time Spectral Amplitude Estimation for Single-Channel Speech Enhancement

Part I explores methods by which to improve the perceptual quality of speech in the presence of acoustic noise. A unified framework for determining short-time spectral amplitude estimators using generalized Gamma *a priori* speech distributions is proposed in Chapter 2. The presence of multiple shape parameters allows flexibility in capturing the statistical behavior of speech. The chapter discusses generalized MAP and MMSE solutions to STSA estimation, and shows special cases to reduce to previous well-known estimators. To supplement the proposed STSA estimators, a soft-decision speech presence uncertainty filter based on GGD priors is discussed. Additionally, channel-specific maximum likelihood estimation of GGD shape parameters is presented. The generalization of the GGD-based STSA framework to include topics such as speech presence

uncertainty and shape parameter estimation serves as a novel contribution.

Next, Chapter 2 proposes a unified framework for designing optimal spectral magnitude estimators assuming equivalent phase of speech and noise components. By assuming phase equivalence, the optimal spectral amplitude estimation problem is effectively projected onto a 1-dimensional subspace of the complex plane. Separate families of estimators assuming generalized gamma distributions (GGDs) for both speech and noise spectral magnitudes are derived, according to the ML, MMSE, or MAP criterion. Solutions are provided in general form, so that estimators can be obtained by substituting statistical shape parameters corresponding to desired speech and noise priors. The stochastic interpretation of spectral subtraction is novel, as are the resulting STSA estimators.

Chapter 3 presents short-time spectral amplitude (STSA) estimation which exploits temporal correlation of spectral speech data. A novel statistical model is proposed for the dynamic behavior of clean speech in the time-frequency spectral domain. The conditional distribution of clean speech STSAs separated by time index is shown to follow a Rician distribution. Furthermore, the conditional distribution of noisy speech based on a neighboring clean spectral components is also shown to follow a Rician model. Using the proposed model of dynamic behavior of clean STSAs, a novel MAP solution to STSA estimation is proposed, which is shown to be a generalized version of that presented in [WG03]. Additionally, a correlation-based SPP filter is derived, which utilizes the model of speech dynamics to incorporate sets of neighboring observed spectral components. The proposed SPP filter is shown to be a generalized version of that presented by Ephraim and Malah in [EM84].

The solutions to single-channel speech enhancement discussed in Chapters 2 and 3 are applied to the Noizeus database, and are generally shown to improve speech quality relative to the MMSE solution from [EM84], in terms of segmental SNR and COSH distance. Informal listening tests are typically consistent with quantitative results.

Finally, Chapter 4 presents a framework for determining improved SPPs using HMM-

based inference. Spectro-temporal data are interpreted as observations from channel-specific two-state models, and HMM-based decoding is applied to estimate true posterior probabilities of active speech. The effectiveness of the proposed SPP framework is assessed by means of pointwise KL distance, spectral distortion, and noise leakage. Though the issue of improved SPPs has been addressed previously, the proposed framework offers a novel approach.

9.2 Front-End Missing Feature Approaches to Noise Robust ASR

Part II addresses the problem of ASR in the presence of background acoustic noise using front-end missing feature methods. Chapter 6.1 presents a novel statistical approach to Mel-domain mask estimation in which reliability measures are derived as conditional probabilities of active speech using a Bayesian approach. Mel-domain power spectra are modeled as χ^2 processes with empirically-determined degrees of freedom. The proposed mask estimation algorithm is applied to the compressive sensing-based MF spectral reconstruction technique from Chapter 7, and is shown to achieve significant improvements in recognition.

Part II offers two novel algorithms for MF spectral reconstruction. The first, presented in Chapter 6, applies HMM-based processing for the estimation of unreliable spectrographic data. Hidden Markov models are utilized to reconstruct corrupted spectral components based on reliable information, unreliable observations, and an underlying signal model, to improve noise robust speech recognition. Separate spectral reconstruction methods are presented to exploit intra-channel (across-time) correlation, inter-channel (across-frequency) correlation, or a combination of both.

The required complexity of HMM processing is reduced by deriving downsampled statistical models. The configurations of such downsampled models are designed through the use of tree-structured quantization, and corresponding statistical parameters are adapted

accordingly. This avoids expensive retraining of codebooks for various resolutions, and reduces memory requirements. One could train a high resolution model, and test using a downsampled model without the need for retraining.

The second algorithm, presented in Chapter 7, exploits the underlying compressibility of speech. Quantitative analysis on the compressibility of spectrographic speech data is provided, motivating the use of minimization of the ℓ_1 -norm in the proposed missing feature estimation technique. This is the first application of compressive sensing theory to front-end noise robustness for ASR, which is not restricted to isolated word recognition.

During experimentation, the proposed spectral reconstruction methods are combined with oracle masks to provide an upper performance bound for our missing feature analysis. Additionally, they are used in series with statistical masks from Chapter 6.1. Both scenarios provide impressive ASR performance relative to the MFCC baseline system.

Finally, Chapter 8 extends previously discussed feature reconstruction to the application of packet loss concealment. Efficient HMM-based estimation techniques for missing speech features are presented, with applications to parametric coding. By assuming features to be observations of hidden Markov processes, the minimum mean-square error solution is derived for estimating unreliable features. Additionally computationally efficient approximations to the derived solutions are explored. When applied to features generally utilized by parametric coding, the proposed estimation methods outperform the baseline repetition scheme in terms of both PSNR and WSD. Though the derivation of the proposed PLC method is similar to previous work, the methods by which to increase computational efficiency provide greatly reduced complexity relative to previous studies, while resulting in negligible degradation.

9.3 Future Work

Future work includes the extension of methods proposed in Part I to multi-channel speech enhancement. With the availability of signals obtained from multiple sensors, the geometric configuration of the sensor set-up can be exploited to determine an accurate estimate of the corruptive noise. This may ultimately lead to improved performance of speech enhancement techniques.

Additionally, spectral estimation techniques proposed in Part I are robust to other signal types, such as music ([OYC06], [WG03]). Particularly, the correlation based enhancement framework can easily be expanded to include 2-dimensional correlation, making it attractive for image denoising ([OYE02], [Jai89]).

Future work also includes the design of spectral reconstruction methods which utilize soft reliability masks. The use of soft masks can be expected to improve system performance, since hard thresholding represents an inherent loss of information. Furthermore, the general compressive sensing framework presented in Chapter 7 can be interpreted as assuming Laplacian *a priori* distributions of speech spectral amplitudes [BMK10]. In future work, various *a priori* distributions can be considered during the reconstruction process.

APPENDIX A

Appendix

A.1 Important Statistical Distributions

A.1.1 The Rayleigh Distribution

The Rayleigh distribution is a continuous, non-negative probability distribution with probability density function [MW71]

$$p(x) = \frac{2x}{\sigma^2} \exp\left(-\frac{x^2}{\sigma^2}\right), \text{ for } x \geq 0 \quad (\text{A.1})$$

Figure A.1 illustrates the Rayleigh distribution for various values of σ^2 . The noncentral moments are expressed as

$$E[x^m] = \sigma^{(m-1)/2} \Gamma\left(\frac{m+2}{2}\right) \quad (\text{A.2})$$

where Γ denotes the Gamma function (see App. A.2.2).

A.1.2 The Rice Distribution

The Rice distribution is a continuous, non-negative, three-parameter probability distribution with pdf [Ric48]

$$p(x) = \frac{x}{\sigma^2} I_0\left(\frac{x\nu}{\sigma^2}\right) \exp\left(-\frac{x^2 + \nu^2}{2\sigma^2}\right), \text{ for } x \geq 0 \quad (\text{A.3})$$

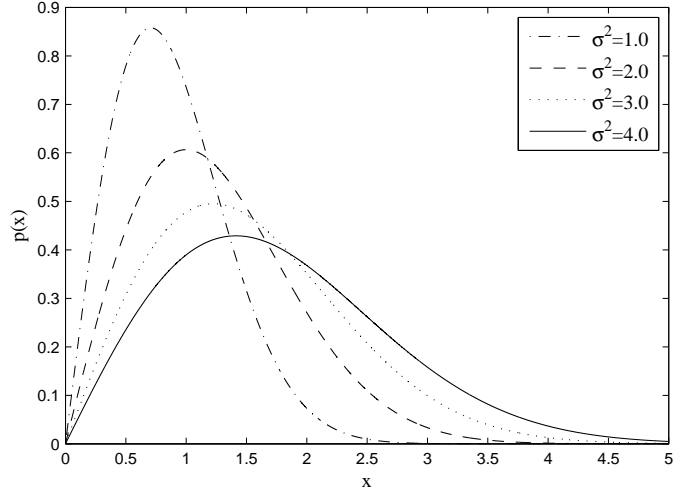


Figure A.1: The Rayleigh distribution for various values of σ^2

Figure A.2 illustrates the Rayleigh distribution for various combinations of σ^2 and ν . The noncentral moments are expressed as

$$E[x^m] = (2\sigma)^m \Gamma\left(1 + \frac{m}{2}\right) L_{m/2}\left(-\frac{\nu^2}{2\sigma^2}\right) \quad (\text{A.4})$$

where L_ν denotes the Laguerre polynomial [AS65]. The first two moments, which occur most regularly, are specified as

$$E[x] = \sqrt{\frac{\pi\sigma^2}{2}} {}_1F_1\left(-\frac{1}{2}; 1; -\frac{\nu^2}{2\sigma^2}\right) \quad (\text{A.5})$$

$$E[x^2] = 2\sigma^2 + \nu^2 \quad (\text{A.6})$$

where ${}_1F_1$ denotes the confluent hypergeometric function (see App. A.2.4).

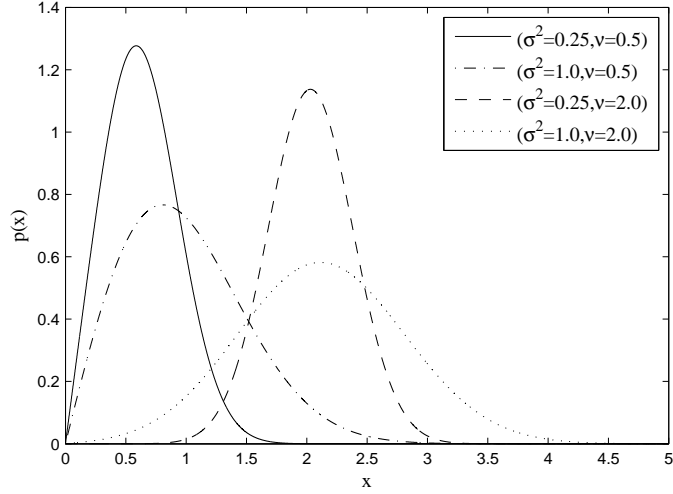


Figure A.2: The Rice distribution for various combinations of σ^2 and ν

A.1.3 The χ^2 Distribution

The χ^2 distribution is a continuous, non-negative, two-parameter probability distribution with pdf [AS65]

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp\left(-\frac{x}{2}\right), \text{ for } x \geq 0 \quad (\text{A.7})$$

where k denotes the degrees of freedom. χ^2 variables commonly occur as the sum of other χ^2 variables with equivalent variances. Figure A.3 illustrates χ^2 distributions for various degrees of freedom k . The noncentral moments are given by

$$E[x^m] = \frac{2^m \Gamma(m + k/2)}{\Gamma(k/2)} \quad (\text{A.8})$$

A.1.4 The Generalized Gamma Distribution

The generalized gamma distribution is a continuous, non-negative, three-parameter family of probability distributions. The GGD has been derived as the optimal distribution for

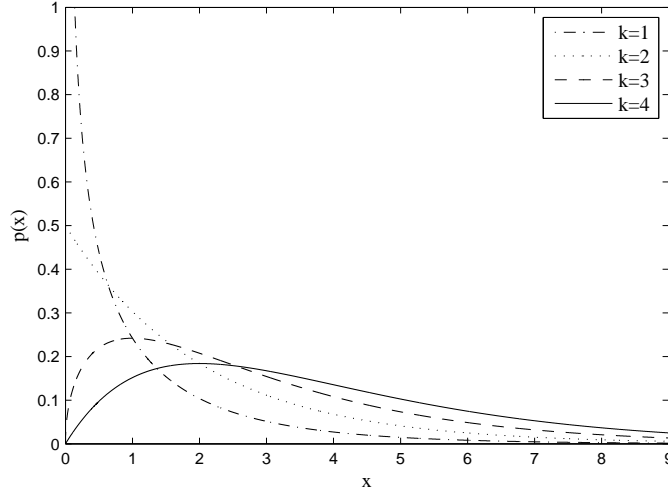


Figure A.3: The χ^2 distribution for various degrees of freedom k

numerous statistical quantities. For example, in [LM67], it is used to model the number of component failures in a given system, as a function of time interval.

The GGD distribution is given by

$$p(x) = \frac{\zeta \beta^\nu}{\Gamma(\nu)} x^{\zeta\nu-1} \exp(-\beta x^\zeta), \text{ for } x \geq 0; \beta, \nu, \zeta > 0 \quad (\text{A.9})$$

Here, ζ and ν serve as shaping parameters, and influence the general behavior of the resulting distribution. Various combinations of shaping parameters result in well-known non-negative distributions. For example, $(\zeta=2, \nu=1)$ provides the Rayleigh distribution and $(\zeta=1, \nu=1)$ leads to the exponential distribution. The scaling parameter β is related to the noncentral second moment.

The noncentral moments are expressed as

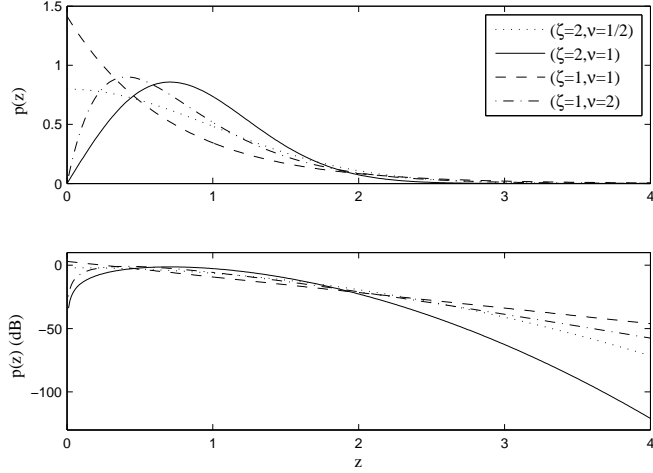


Figure A.4: The generalized gamma distribution for various shaping parameter pairs: For illustrative purposes, the variance was normalized to unity.

$$E[x^m] = \frac{\prod_{i=0}^{m-1} (\nu + i)}{\beta^m} \text{ for } \zeta = 1 \quad (\text{A.10})$$

$$E[x^m] = \frac{\Gamma(\nu + m/2)}{\Gamma(\nu) \beta^{m/2}} \text{ for } \zeta = 2 \quad (\text{A.11})$$

Figure A.4 provides probability distribution functions (pdfs) of the GGD for various shaping parameter pairs. Specifically, the top panel shows $p(z)$ in the linear scale, illustrating the effect of shaping parameters on GGDs. The bottom panel shows $p(z)$ in the log scale, illustrating the relative prominence of distribution tails for each set of shaping parameters. For illustrative purposes, the variance is normalized to unity in each case.

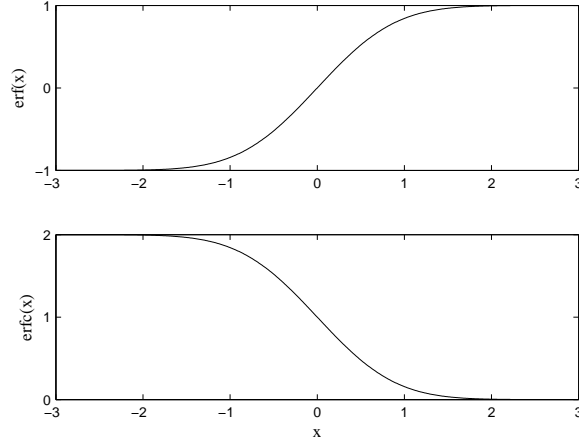


Figure A.5: The erf and erfc functions

A.2 Special Functions

A.2.1 The Gauss Error and Gauss Complementary Error Functions

The Gauss error function, denoted by erf , is derived from the cumulative distribution function of a zero-mean, normal distribution with variance $1/2$ [AS65]

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-\tau^2) d\tau \quad (\text{A.12})$$

The related complementary error function, denoted by erfc , is defined as

$$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-\tau^2) d\tau \quad (\text{A.13})$$

Figure A.5 illustrates the erf and erfc functions. The Taylor series expansion of the Gauss error function is given by [AS65]

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{x! (2x+1)} \quad (\text{A.14})$$

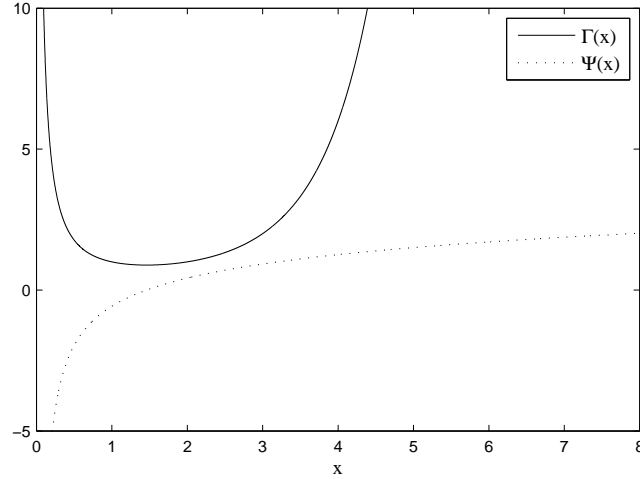


Figure A.6: The Gamma (Γ) and Digamma (Ψ) functions

A.2.2 The Gamma and Digamma Functions

The Gamma function is given by [Arf85]

$$\Gamma(x) = \int_0^{\infty} t^{x-1} \exp(-t) dt \quad (\text{A.15})$$

which reduces for $x \in \mathbb{R}$

$$\Gamma(x) = (x-1)!, \text{ for } x \in \mathbb{R}, x > 0 \quad (\text{A.16})$$

The Digamma function arises as the derivative of Γ in the log domain, and is expressed as [AS65]

$$\Psi(x) = \frac{d/dx \Gamma(x)}{\Gamma(x)} \quad (\text{A.17})$$

Figure A.6 illustrates the Gamma (Γ) and Digamma (Ψ) functions.

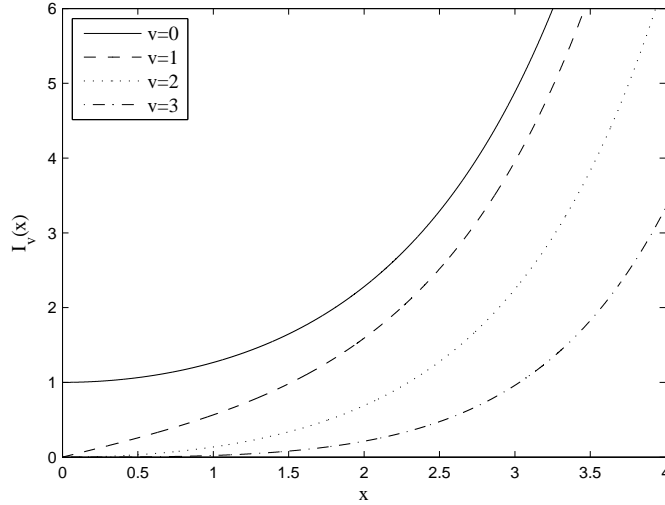


Figure A.7: The modified Bessel function of the first kind for various orders v

A.2.3 The Modified Bessel Function of the First Kind

The modified Bessel function of the first kind of the v^{th} -order is defined for $x \in \mathbb{R}$ as [AS65]

$$I_v(x) = \frac{1}{\pi} \int_0^\pi \exp(x \cos \theta) \cos(v\theta) d\theta \quad (\text{A.18})$$

which can be expressed as an infinite sum

$$I_v(x) = \left(\frac{x}{2}\right)^v \sum_0^\infty \frac{(x^2/4)^k}{k! \Gamma(v+k+1)} \quad (\text{A.19})$$

The following large-value approximation of the 0^{th} -order function from [AS65] is commonly applied in speech enhancement studies

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}} \exp(x) \quad (\text{A.20})$$

Figure A.7 illustrates $I_v(x)$ for various orders v .

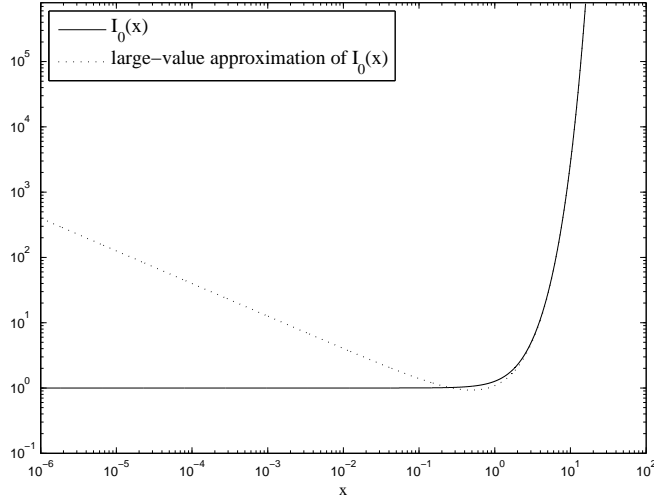


Figure A.8: The 0^{th} -order modified Bessel function of the first kind, and its large-value approximation from Eq. A.20

Figure A.7 illustrates $I_0(x)$ and its large-value approximation from Eq. A.20.

A.2.4 The Confluent Hypergeometric Function

The confluent hypergeometric function appears as solution within several statistical speech enhancement problems. It is expressed in integral form as [Mac48]

$${}_1F_1(\nu; b; x) = \frac{\Gamma(b)}{\Gamma(\nu)\Gamma(b-\nu)} \int_0^1 \exp(xu) u^{\nu-1} (1-u)^{b-\nu-1} du \quad (\text{A.21})$$

It is interesting to note the role of ${}_1F_1$ in the odd noncentral moments of the Rician distribution. Figure A.9 illustrates the confluent hypergeometric function for $b=1$ and for various values of ν .

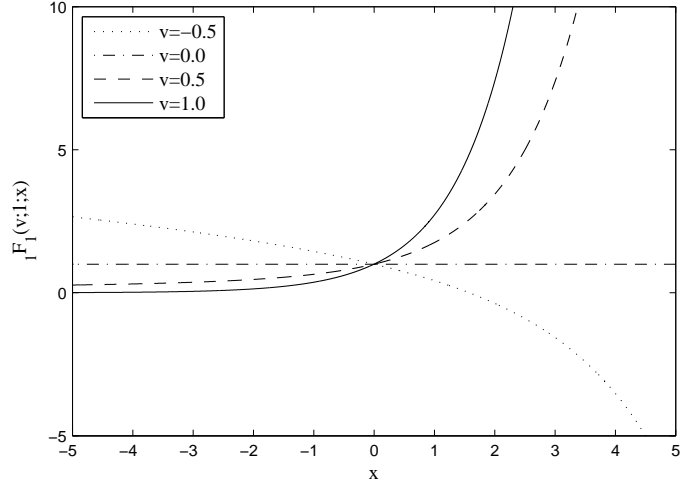


Figure A.9: The confluent hypergeometric function for $b=1$ and for various values of ν

A.2.5 The Parabolic Cylinder Function

The v^{th} -order parabolic cylinder function can be expressed in terms of ${}_1F_1$ [SO87]

$$D_v(x) = 2^{x/2} \exp(-x^2/4) {}_1F_1(-v/2; 1/2; x^2/2) \quad (\text{A.22})$$

or in integral form

$$D_v(x) = \frac{2^{-v/2-1} x^v}{\Gamma(-v)} \int_0^\infty t^{-v/2-1} \exp\left(\frac{-t}{x^2} - \sqrt{2t}\right) dt \quad (\text{A.23})$$

Figure A.10 illustrates the parabolic cylinder function for various orders v .

A.3 Derivation of Eq. 2.30

Using a geometric approach to spectral subtraction, and applying the law of cosines leads to

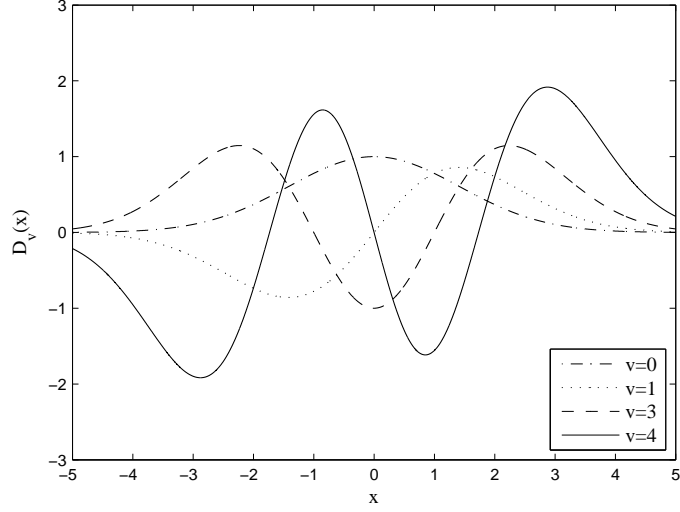


Figure A.10: The parabolic cylinder function for various orders v

$$R_k^2 = A_k^2 + D_k^2 - 2A_k D_k \cos \theta_k \quad (\text{A.24})$$

Solving for A_k using the quadratic equation results in

$$A_k = D_k \cos \theta_k + \sqrt{D_k^2 \cos^2 \theta_k - D_k^2 + R_k^2} \quad (\text{A.25})$$

Grouping terms in Eq. A.25 to match those in Eq. 2.28 yields

$$A_k = R_k - \left(\sqrt{\gamma_k} - \sqrt{\gamma_k - \sin^2 \theta_k} + \cos \theta_k \right) D_k \quad (\text{A.26})$$

A.4 Derivation of the Identity in Eq. 2.49

Define the integral I_n as

$$I_n = \int x^n e^{cx} dx. \quad (\text{A.27})$$

This can be expressed recursively by

$$I_n = \begin{cases} \frac{1}{c}x^n e^{cx} - \frac{n}{c} \int x^{n-1} e^{cx} dx, & \text{if } n > 0 \\ \frac{1}{c}e^{cx}, & \text{if } n = 0 \end{cases} \quad (\text{A.28})$$

I_n can be expanded into

$$I_n = \frac{1}{c}x^n e^{cx} - \frac{n}{c^2}x^{n-1}e^{cx} + \frac{n(n-1)}{c^3}x^{n-2}e^{cx} - \dots + \frac{n!}{c^{n+1}}e^{cx}, \quad (\text{A.29})$$

which can be grouped concisely into a summation

$$I_n = \frac{e^{cx}}{c} \sum_{k=0}^n \left[(-1)^k \frac{n!}{(n-k)!} \frac{x^{n-k}}{c^k} \right]. \quad (\text{A.30})$$

A.5 Evaluating the Indeterminant Form of Eq. 2.50

The EMMSE gain function is given by (Eq. 2.50). It can be observed that

$$G_{EMMSE}^{\nu_x}(\xi_k, \gamma_k) \Big|_{\mu_k=0} = \frac{0}{0}, \quad (\text{A.31})$$

which is an indeterminant form. We therefore apply L'Hoptial's Rule to evaluate the gain at $\mu_k=0$. Let $N(\mu_k)$ and $D(\mu_k)$ refer to the numerator and denominator of $G_{EMMSE}^{\nu_x}(\xi_k, \gamma_k)$, respectively, scaled by $\exp(\mu_k)$

$$N(\mu_k) = \exp(\mu_k) (-1)^{\nu_x+1} \nu_x! + \sum_{k=0}^{\nu_x} \left((-1)^k \frac{\nu_x!}{(\nu_x-k)!} \mu_k^{\nu_x-k} \right) \quad (\text{A.32})$$

$$D(\mu_k) = \mu_k \exp(\mu_k) (-1)^{\nu_x} (\nu_x-1)! + \sum_{k=0}^{\nu_x-1} \left((-1)^k \frac{(\nu_x-1)!}{(\nu_x-k-1)!} \mu_k^{\nu_x-k} \right)$$

L'Hopital's Rule states that

$$G_{EMMSE}^{\nu_x}(\xi_k, \gamma_k) \Big|_{\mu_k=0} = \frac{\frac{\partial^m}{\partial \mu_k^m} N(\mu_k) \Big|_{\mu_k=0}}{\frac{\partial^m}{\partial \mu_k^m} D(\mu_k) \Big|_{\mu_k=0}}, \text{ for } m \geq 0. \quad (\text{A.33})$$

Note that the last term in $N(\mu_k)$ is simply a summation of scaled powers of μ_k , with the highest power being ν_x . Thus, in the $(\nu_x + 1)^{th}$ derivative of $N(\mu_k)$ with respect to μ_k , the summation will disappear

$$\frac{\partial^{\nu_x+1}}{\partial \mu_k^{\nu_x+1}} N(\mu_k) = (-1)^{2\nu_x+2} \nu_x! \exp(\mu_k). \quad (\text{A.34})$$

Likewise, the last term in $D(\mu_k)$ is a linear function of μ_k with highest power ν_x

$$\frac{\partial^{\nu_x+1}}{\partial \mu_k^{\nu_x+1}} D(\mu_k) = (-1)^{2\nu_x+1} (\nu_x - 1)! (\mu_k \exp(\mu_k) - (\nu_x + 1) \exp(\mu_k)). \quad (\text{A.35})$$

The indeterminant form of $G_{EMMSE}^{\nu_x}(\xi, \gamma)$ can then be evaluated as

$$\begin{aligned} & G_{EMMSE}^{\nu_x}(\xi, \gamma) \Big|_{\mu_k=0} \quad (\text{A.36}) \\ &= \frac{[(-1)^{2\nu_x+2} \nu_x! \exp(\mu_k)]_{\mu_k=0}}{[(-1)^{2\nu_x+1} (\nu_x - 1)! (\mu_k \exp(\mu_k) - (\nu_x + 1) \exp(\mu_k)) \cdot]_{\mu_k=0}} = \frac{\nu_x}{\nu_x + 1}. \end{aligned}$$

REFERENCES

- [Ace93] A. Acero. *Acoustic and Environmental Robustness in Automatic Speech Recognition*. Kluwer, 1993.
- [Arf85] G. Arfken. *Mathematical Methods for Physicists*. Academic Press, 1985.
- [AS65] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1965.
- [BA02] A. Bernard and A. Alwan. “Low-bitrate Distributed Speech Recognition for Packet-based and Wireless Communication.” *IEEE Transactions on Speech and Audio Processing*, **10**(8):570–580, 2002.
- [BA08a] B. J. Borgstrom and A. Alwan. “HMM-Based Estimation of Unreliable Spectral Components for Noise Robust Speech Recognition.” *Interspeech*, pp. 1769–1772, 2008.
- [BA08b] B. J. Borgstrom and A. Alwan. “A Low Complexity Parabolic Lip Contour Model With Speaker Normalization For High-Level Feature Extraction in Noise Robust Audio-Visual Speech Recognition.” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, **38**(6):1273–1280, 2008.
- [BA09] B. J. Borgstrom and A. Alwan. “Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR.” *IEEE Signal Processing Letters*, **16**(5):398–401, 2009.
- [BA10a] B. J. Borgstrom and A. Alwan. “HMM-Based Reconstruction of Unreliable Spectrographic Data for Noise Robust Speech Recognition.” *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.
- [BA10b] B. J. Borgstrom and A. Alwan. “Improved Speech Presence Probabilities Using HMM-Based Inference, with Applications to Speech Enhancement and ASR.” *IEEE Journal of Selected Topics in Signal Processing*, 2010.
- [BJC00] J. Barker, L. Josifovski, M. Cooke, and P. Green. “Soft Decisions in Missing Feature Data Techniques for Robust Automatic Speech Recognition.” *ICSLP*, pp. 373–376, 2000.
- [BMK10] S. D. Babacan, R. Molina, and A. K. Katsaggelos. “Bayesian Compressive Sensing Using Laplace Priors.” *IEEE Trans. on Image Processing*, **19**(1):53–63, 2010.

- [Bol79] S. F. Boll. “Suppression of Acoustic Noise in Speech Using Spectral Subtraction.” *IEEE Trans, on Acoustics, Speech, and Signal Processing*, **27**:113–120, 1979.
- [BSM79] M. Berouti, M. Schwartz, and J. Makhoul. “Enhancement of speech corrupted by acoustic noise.” *IEEE ICASSP*, pp. 208–211, 1979.
- [BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [CBA03] X. Cui, A. Bernard, and A. Alwan. “A Noise-Robust ASR Back-end Technique Based on Weighted Viterbi Recognition.” *Eurospeech*, pp. 2169–2172, 2003.
- [CGJ01] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. “Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data.” *Speech Communication*, **34**:267–285, 2001.
- [CL07] B. Chen and P. C. Loizou. “A Laplacian-based MMSE Estimator for Speech Enhancement.” *Speech Communication*, **49**:134–143, 2007.
- [CMG97] M. P. Cooke, A. Morris, and P. D. Green. “Missing Data Techniques for Robust Speech Recognition.” *ICASSP*, **2**:863–866, 1997.
- [Coh] “<http://webee.technion.ac.il/Sites/People/IsraelCohen/>”.
- [Coh02] I. Cohen. “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator.” *IEEE Signal Processing Letters*, **9**(4):113–116, 2002.
- [Coh04] I. Cohen. “Modeling speech signals in the timefrequency domain using GARCH.” *Signal Processing*, **84**:2453–2459, 2004.
- [Coh05] I. Cohen. “Speech Enhancement Using Super-Gaussian Speech Models and Noncausal a Priori SNR Estimation.” *Speech Communication*, **47**:336–350, 2005.
- [CT91] T. M. Cover and J. M. Thomas. *Elements of Information Theory*. Kluwer Inc., 1991.
- [CW69] S. C. Choi and R. Wette. “Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias.” *Technometrics*, **11**(4):683–690, 1969.
- [CW08] E. J. Candes and M. B. Wakin. “An Introduction to Compressive Sampling.” *IEEE Signal Processing Magazine*, **25**(2):21–30, 2008.

- [DAP00] L. Deng, A. Acero, M. Plumpe, and X. Huang. “Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments.” *ICSLP*, pp. 806–809, 2000.
- [Doc07] ETSI Standard Doc. “Speech Processing, Transmission, and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithms; Compression Algorithms.” *ETSI ES 202 050 v1.1.1 (2007-10)*, 2007.
- [Don06] D. L. Donoho. “Compressed Sensing.” *IEEE Trans. on Information Theory*, **52**(4):1289–1306, 2006.
- [DTI05] T. H. Dat, K. Takeda, and F. Itakura. “Generalized Gamma Modeling of Speech and Its Online Estimation for Speech Enhancement.” *ICASSP*, **4**:181–184, 2005.
- [EHH07] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. “Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors.” *IEEE Trans. Audio, Speech, and Language Processing*, **15**(6):1741–1752, 2007.
- [EM84] Y. Ephraim and D. Malah. “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator.” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **32**(6):1109–1121, 1984.
- [EM85] Y. Ephraim and D. Malah. “Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator.” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **33**(2):443–445, 1985.
- [Esp68] R. Esposito. “On a relation between detection and estimation in decision theory.” *Information and Control*, **12**(2):116–120, 1968.
- [FV01] T. Fingscheidt and P. Vary. “Softbit Speech Decoding: a New Approach to Error Concealment.” *IEEE Trans. on Speech and Audio Processing*, **9**(3):240–251, 2001.
- [GBC01] P. Green, J. Barker, M. Cooke, and L. Josifovski. “Handling Missing and Unreliable Information in Speech Recognition.” *Proc. AISTATS*, 2001.
- [GBM08] T. Gerkmann, C. Breithaupt, and R. Martin. “Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors.” *IEEE Trans. on Audio, Speech, and Language Processing*, **16**(5):910–919, 2008.
- [GG01] M. Gupta and A. Gilbert. “Robust Speech Recognition using Wavelet Coefficient Features.” *Speech Recognition and Understanding Workshop*, 2001.

- [GLJ93] J. S. Garofolo, L. F. Lamel, W. M. Fisher and J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. "TIMIT Acoustic-Phonetic Continuous Speech Corpus." *Linguistic Data Consortium*, 1993.
- [Gon95] Y. Gong. "Speech recognition in noisy environments: a survey." *Speech Communication*, **16**(3):261–291, 1995.
- [Har75] J. A. Hartiga. *Clustering Algorithms*. Wiley, 1975.
- [Haz06] T. J. Hazen. "Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition." *IEEE Trans. on Speech and Audio Processing*, **14**(3):1082–1089, 2006.
- [Her90] H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech." *Journal of the Acoustic Society of America*, **87**(4):1738–1752, 1990.
- [HL07] Y. Hu and P. Loizou. "Subjective evaluation and comparison of speech enhancement algorithms." *Speech Communication*, **49**:588–601, 2007.
- [HWh07] K. Hermus, P. Wambacq, and H. Van hamme. "A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition." *EURASIP Journal on Advances in Signal Processing*, **2007**:1–15, 2007.
- [IH06] V. Ion and R. Haeb-Umbach. "Uncertainty Decoding for Distributed Speech Recognition Over Error-Prone Networks." *Speech Communication*, **48**:1435–1446, 2006.
- [IH07] V. Ion and R. Haeb-Umbach. "Multi-Resolution Soft-Features for Channel-Robust Distributed Speech Recognition." *Interspeech*, pp. 594–597, 2007.
- [Jai89] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.
- [JM76] A. H. Gray Jr. and J. D. Markel. "Distance Measures for Speech Processing." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **24**(5):380–391, 1976.
- [KH09] W. Kim and J. H. L. Hansen. "Time-Frequency Correlation Based Missing-Feature Reconstruction for Robust Speech Recognition in Band-Restricted Conditions." *IEEE Transactions on Audio, Speech, and Language Processing*, **17**(7):1292–1304, 2009.
- [KH10] W. Kim and J. H. L. Hansen. "Missing-Feature Reconstruction by Leveraging Temporal Spectral Correlation for Robust Speech Recognition in Background Noise Conditions." *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

- [Kon06] A. M. Kondo. *Digital Speech*, chapter Sampling and Quantization, pp. 46–48. Wiley Inc., 2006.
- [KS06] W. Kim and R. M. Stern. “Band-Independent Mask Estimation for Missing-Feature Reconstruction in the Presence of Unknown Background Noise.” In *Proceedings of ICASSP*, volume 1, pp. 305–308, 2006.
- [LB92] P. Lockwood and J. Boudy. “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projections, for robust recognition in cars.” *Speech Communication*, **11**(2-3):215–228, 1992.
- [LL08] Y. Lu and P. Loizou. “A geometric approach to spectral subtraction.” *Speech Communication*, **50**:453–466, 2008.
- [LM67] J. H. Lienhard and P. L. Meyer. “A Physical Basis for the Generalized Gamma Distribution.” *Quarterly of Applied Mathematics*, **25**:330–334, 1967.
- [Loi05] P. Loizou. “Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum.” *IEEE Trans. on Speech and Audio Processing*, **13**(5):857–869, 2005.
- [Loi07] P. C. Loizou. *Speech Enhancement: Theory and Practice*. Taylor and Francis, 2007.
- [LV03] T. Lotter and P. Vary. “Noise Reduction by Maximum a Posterior Spectral Amplitude Estimation with Supergaussian Speech Modeling.” *International Workshop on Acoustic Echo and Noise Control*, **1**:83–86, 2003.
- [LV05] T. Lotter and P. Vary. “Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model.” *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, 2005.
- [Mac48] A. D. MacDonald. “Properties of the Confluent Hypergeometric Function.” *MIT Technical Report*, (84), 1948.
- [Mar01] R. Martin. “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics.” *IEEE Trans. Speech and Audio Processing*, **9**(5):504–512, 2001.
- [Mar05] R. Martin. “Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors.” *IEEE Trans. Speech and Audio Processing*, **13**(5):845–856, 2005.
- [MM80] R. J. McAulay and M. L. Malpass. “Speech Enhancement Using a Soft-Decision Noise Suppression Filter.” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, **28**(2):137–145, 1980.

- [MW71] R. N. McDonough and A. D. Whalen. *Detection of Signals in Noise*. Academic Press, 1971.
- [OYC06] S.-H. Oh, W.-J. Yoan, Y.-H. Cho, and K.-S. Park. “A New Spectral Enhancement Algorithm in MP3 Audio.” *ICCE*, pp. 285–286, 2006.
- [OYE02] R. Okten, L. Yaroslavsky, K. Egiazarian, and J. Astola. “Transform Domain Approaches for Image Denoising.” *Journal of Electronic Imaging*, **11**(2):149–156, 2002.
- [P800] ITU P.862. “Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.” 2000.
- [PA93] K. K. Paliwal and B. S. Atal. “Efficient vector quantization of LPC parameters at 24 bits/frame.” *IEEE Trans. on Speech and Audio Processing*, **1**(1):3–14, 1993.
- [Pal99] K. K. Paliwal. “Decorrelated and lifted filter-bank energies for robust speech recognition.” *EUROPSEECH*, pp. 85–88, 1999.
- [PC08] E. Plourde and B. Champaign. “Auditory-Based Spectral Amplitude Estimators for Speech Enhancement.” *IEEE Trans. on Speech and Audio Processing*, **52**(7):1614–1623, 2008.
- [Pea00] D. Pearce. “Enabling New Speech Driven Services for Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends.” In *Proceedings of the Speech Applications Conference (AVIOS)*, volume 5, pp. 1–5, 2000.
- [PSP02] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, J. C. Segura, and J. Rubio. “HMM-Based Methods for Channel Error Mitigation in Distributed Speech Recognition.” *ICSLP*, pp. 2707–2710, 2002.
- [PSP03] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, and A. Torre. “HMM-Based Channel Error Mitigation and its Applications to Distributed Speech Recognition.” *Speech Communication*, **41**:549–561, 2003.
- [PSS01] A. M. Peinado, V. Sanchez, J. C. Segura, and J. L. Perez-Cordoba. “MMSE-Based Channel Mitigation for Distributed Speech Recognition.” *Eurospeech*, pp. 2205–2208, 2001.
- [PSS04] R. Prasadi, H. Saruwatari, and K. Shikano. *Independent Component Analysis and Blind Signal Separation, Chp. 1: Single Channel Speech Enhancement: MAP Estimation Using GGD Prior Under Blind Setup*. Springer, 2004.

- [Qua01] T. F. Quatieri. *Discrete-time Speech Signal Processing*. Prentice Hall, 2001.
- [Rab89] L. R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.” In *Proceedings of the IEEE*, volume 77, pp. 257–286, 1989.
- [Ram00] B. R. Ramakrishnan. “Reconstruction of Incomplete Spectrograms for Robust Speech Recognition.” *PhD. Dissertation, CMU*, 2000.
- [RH93] L. Rabiner and B. H. Huang. *Fundamentals of Speech Recognition*, chapter Theory and Implementation of Hidden Markov Models, pp. 334–339. Prentice Hall Inc., 1993.
- [Ric48] S. O. Rice. “Statistical Properties of a Sine Wave Plus Random Noise.” *Bell System Technical Journal*, **27**:109–157, 1948.
- [Ril89] M. D. Riley. *Speech time-frequency representations*. Springer, 1989.
- [RMA06] C. A. Rodbro, M. N. Murthi, S. V. Anderson, and S. H. Jensen. “Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP.” *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(5):1609–1623, 2006.
- [Rom08] J. Romberg. “Imaging via Compressive Sensing.” *IEEE Signal Processing Magazine*, **25**(2):14–20, 2008.
- [RS05] B. Raj and R. Stern. “Missing Feature Approaches in Speech Recognition.” *IEEE Signal Processing Magazine*, pp. 101–116, 2005.
- [RSS04] B. Raj, M. L. Seltzer, and R. M. Stern. “Reconstruction of Missing Features for Robust Speech Recognition.” *Speech Communication*, **43**:275–296, 2004.
- [SA97] B. Stroppe and A. Alwan. “A Model of Dynamic Auditory Perception and its Application to Robust Word Recognition.” *IEEE Trans. on Speech and Audio Processing*, **5**(5):451–464, 1997.
- [Sh04] M. V. Segbroeck and H. V. Hamme. “Vector-Quantization based Mask Estimation for Missing Data Automatic Speech Recognition.” *ICSLP*, 2004.
- [SKS99] J. Sohn, N. S. Kim, and W. Sung. “A statistical model-based VAD.” *IEEE Signal Processing Letters*, **16**(1):1–3, 1999.
- [SO87] J. Spanier and K. B. Oldham. *An Atlas of Functions*. Hemisphere, 1987.
- [SRS04] M. L. Seltzer, B. Raj, and R. M. Stern. “A Bayesian Classifier for Spectrographic Mask Estimation for Missing Feature Speech Recognition.” *Speech Communication*, **43**:379–393, 2004.

- [SS77] H. Solomon and M. A. Stephens. "Distribution of a Sum of Weighted Chi-Square Variables." *American Statistical Association*, **72**(360):881–885, 1977.
- [TJ00] C. Tellambura and D. S. Jayalath. "Generation of Bivariate Rayleigh and Nakagami-m Fading Envelopes." *IEEE Communication Letters*, **4**(5):170–172, 2000.
- [VGC99] A. Vizinho, P. Green, M. Cooke, and L. Josifovski. "Missing Data Theory, Spectral Subtraction And Signal-To-Noise Estimation For Robust Asr: An Integrated Study." *Eurospeech*, pp. 2407–2410, 1999.
- [VL98] O. Viikki and K. Laurila. "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition." *Speech Communication*, **25**:133–147, 1998.
- [VS93] A. Varga and H. J. M. Steeneken. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems." *Speech Communication*, **12**(3):247–251, 1993.
- [WG03] P. J. Wolfe and S. J. Godsil. "Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement." *EURASIP J. Appl. Signal Processing*, **10**:1043–1051, 2003.
- [Wie49] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.
- [WL82] D. L. Wang and J. S. Lim. "The Unimportance of Phase in Speech Enhancement." *IEEE Trans. Acoustics, Speech, and Signal Processing*, **30**(4):679–681, 1982.
- [YKO] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. "The HTK Book."
- [ZA03] Q. Zhu and A. Alwan. "Nonlinear Feature Extraction for Robust Recognition in Stationary and Non-Stationary Noise." *Computer, Speech, and Language*, **14**(4):381–402, 2003.