Speech enhancement methods based on masking properties

You, Chang Huai

2006

You, C. H. (2006). Speech enhancement methods based on masking properties. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/3908>

<https://doi.org/10.32657/10356/3908>

Nanyang Technological University

Downloaded on 26 Apr 2023 08:26:51 SGT

Speech Enhancement Methods Based on Masking Properties

YOU CHANG HUAI

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2005

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Professor Koh Soo Ngee, and my co-supervisor, Associate Professor Rahardja Susanto, for their invaluable guidance and pertinent advices during my study. Moreover, I would like to thank their constant encouragement, their enormous time and effort that they have made pertaining to this research and thesis writing.

I am deeply grateful to my parents, Mr. You Bao Hui and Ms. Zheng Hui Yu, and my wife, Jiang Ping, and my two sons, Li Ming and Li Hong, for their endless love, support and encouragement.

I would also like to thank my final year project students from National University of Singapore, Ms. Chen Kexin, Mr. Tan Chun Wee and Ms. Sem Pak Ling, for sharing with me their projects and many interesting and helpful discussions.

Finally, I am grateful to the enthusiasm and help from the staff members, Dr. Louis Shue, Dr. Zhang Jian Min and Mr. Li Chang Chun of the Institute for Infocomm Research. I would also like to thank those who have ever helped to do the listening tests.

Summary

This thesis focuses on the development of robust speech enhancement methods based on masking properties of the human auditory system. First, motivation of the work is stated, followed by an overview of the history of speech enhancement research. Some of the important speech enhancement algorithms are outlined and masking properties of the human auditory system are reviewed. A modified relative threshold offset scheme for the approximation of masking threshold is then proposed.

With the realization of the strengths of the Ephraim & Malah (E-M) minimum mean-square error (MMSE) spectral suppression method that is based on the non-linearity of the averaging procedure, the effects of different orders of the MMSE method are investigated, which is named β -order MMSE. It is noticed that the advantage of the musical tone elimination of E-M MMSE can be extended to β -order MMSE estimation. Subsequently three adaptive β -order MMSE methods are proposed. The first method considers the adaption of β value based on frame SNR which is defined as the SNR over a frame. The second method takes into account speech-presence uncertainty and time-frequency adaptation of β value by introducing the definition of local signal-to-noise ratio (SNR). The third method incorporates the perceptual properties of the human auditory system into an adaptive β -order short-time spectral amplitude (STSA) MMSE speech enhancement system. It basically determines the value of β using the masking threshold derived from a preliminary version of enhanced speech. It is found that the three proposed speech enhancement methods have good tradeoff for minimizing both the levels of speech distortions and residual noise. Furthermore, the noise elimination effect of β -order MMSE estimation is discussed, and a simplified scheme to reduce the computational complexity of the β -order MMSE speech enhancement algorithm is developed.

In order to exploit effectively the human perceptual properties in a Kalman filtering speech enhancement system, a masking-based Kalman filtering method through subband analysis is

proposed. The noise variance estimate used in the Kalman filtering process in each subband is modified according to the masking threshold mapped to the subband. As the method involves masking properties, subband decomposition and Kalman filtering, the contribution from each of them to the performance of the speech enhancement scheme is investigated through computer simulations. Both white noise model and colored noise model are examined for different noise cases.

In the study of subspace speech enhancement techniques, a subspace-based audible noise reduction method is proposed. In the method, the masking properties are incorporated into a subspace scheme. This method emphasizes the suppression of audible noise in the eigendomain, and the gain of the eigen-filter is regulated by an audible noise reduction criterion. The eigendomain audible noise reduction algorithm is developed for white noise through defining the concept of masking-eigenvalue. Then the algorithm is extended to the colored noise case. Furthermore, in order to develop a more effective masking-based subspace enhancement system, we propose an invertible frequency eigendomain transformation. The implementation of both white and colored noise models is further studied.

The proposed speech enhancement methods are compared with many well-known enhancement methods in terms of various objective measures such as segmental SNR, PESQ, Itakura Saito distortion and modified Bark spectral distortion, as well as subjective measures including mean opinion score (MOS), informal listening test and spectrogram inspection. Through a large number of simulations, we verify the effectiveness and significance of the proposed methods.

According to the attributes of each proposed speech enhancement methods, two proposed speech enhancement algorithms are chosen for some practical applications. The first application is to apply speech enhancement in an interior car noise environment, where the masking-based β -order MMSE scheme is further modified according to the general characteristics of car noise. The reduction of noise with the modified scheme is verified to be significant and effective by simulations using real-life recorded car noise at various car speeds and different scenarios. In addition, the proposed masking-based Kalman filtering speech enhancement method is applied as

a preprocessing stage in a speech recognition system. With a real-life recorded speech database and a commercial speech recognition system, it is found that the proposed Kalman filtering enhancement method is effective in improving the speech recognition rate.

Contents

Summary	i
Table of Contents	iv
List of Abbreviations and Symbols	viii
List of Figures	xvi
List of Tables	xvi
1 Introduction	1
1.1 Motivation	1
1.2 Scope of the Thesis	3
1.3 Contributions	5
2 Background of Masking-Based Speech Enhancement System	6
2.1 A Brief History of Speech Enhancement	6
2.2 Approaches of Speech Enhancement	9
2.2.1 The Speech Production Model	12
2.2.2 The Signal Characteristics Model	14
2.2.3 The Speech Receiver Model	18
2.3 Masking Properties of Human Auditory System	20
2.3.1 Masking Threshold	20
2.3.2 Modified Relative Threshold Offset	21
2.4 Methodology of Performance Evaluation	24
2.4.1 Subjective Measure	25
2.4.2 Objective Measure	26

2.5	Database for Performance Assessment	29
3	Spectral Suppression Method	31
3.1	Introduction	31
3.2	Proposed β -Order MMSE Short-Time Spectral Suppression	34
3.2.1	Gain Function of β -Order MMSE Estimator	34
3.2.2	Characteristic of β -Order STSA MMSE Estimator	37
3.2.3	Investigation of MMSE Estimator for Constant β Value	40
3.3	Adaptive β -Order MMSE	42
3.4	A Time-Frequency Adaptive β -Order MMSE Algorithm with Speech-Presence Uncertainty	46
3.4.1	Time-Frequency Adaption of β	46
3.4.2	Suppression Gain Incorporating Speech-Presence Uncertainty	48
3.5	Proposed β -Order MMSE with Masking Properties	50
3.5.1	Gain of a Masking-Based β -Order MMSE Method	50
3.5.2	Adaptation of β Value Based on Masking Properties	51
3.5.3	Implementation of Masking-based Speech Enhancement Method	54
3.6	Analysis of Musical Tone Elimination	54
3.7	Reduction of Computational Complexity	57
3.8	Performance Evaluation	58
3.9	Summary	68
4	Kalman Filtering Method	70
4.1	Introduction	70
4.2	Kalman Filtering for Speech Enhancement	72
4.2.1	White Noise Model for Speech Enhancement	72
4.2.2	Colored Noise Model for Speech Enhancement	74
4.3	Incorporating Masking Properties into Subband Kalman Filtering	75
4.3.1	Subband Filter	75

4.3.2	AR Parameters Estimation in a Subband	76
4.3.3	Incorporating Masking into Subband Kalman Filtering	79
4.4	System Configuration of Proposed Scheme	87
4.4.1	Determination of the Number of Subbands	87
4.4.2	Selection of AR Order	87
4.4.3	Verification of Various Functions	88
4.4.4	Monitoring of the Effect in Each Iteration	89
4.5	Performance Evaluation	90
4.6	Summary	97
5	Subspace Method	98
5.1	Introduction	98
5.2	Subspace Speech Enhancement	100
5.3	Proposed Invertible Frequency Eigendomain Transformation	101
5.3.1	Proposed Subspace-Based Audible Noise Reduction	106
5.3.2	Proposed Approach for Colored Noise	112
5.3.3	Implementation	114
5.4	Performance Evaluation	116
5.5	Summary	126
6	Performance Comparison among the Proposed Speech Enhancement Methods	128
6.1	Introduction	128
6.2	Comparison among the Enhancement Methods	129
6.2.1	Characteristics of the Enhancement Methods	130
6.2.2	Performance Comparison	131
6.3	Summary	133
7	Speech Enhancement Applications	141
7.1	Speech Enhancement for Car Mobile Communication	141
7.1.1	Modified Adaptation of β Value	142

7.1.2	Performance Evaluation	145
7.2	Speech Enhancement for Speech Recognition	146
7.2.1	Introduction of Speech Recognition System	157
7.2.2	Speech Enhancement for Speech Recognition	159
7.3	Summary	162
8	Conclusion and Recommendations	164
8.1	Conclusion	164
8.2	Recommendations for Further Research	167
	Author's Publications	170
	Bibliography	172

List of Abbreviations and Symbols

Abbreviations

ANR	Audible noise reduction
AR	Autoregressive
CFE-ANR-1	ANR with conventional FET and white noise model
CFE-ANR-2	ANR with conventional FET and colored noise model
DFT	Discrete Fourier transform
FET	Frequency eigendomain transformation
HMM	Hidden Markov model
IDFT	Inverse discrete Fourier transform
IFE-ANR-1	ANR with invertible FET and white noise model
IFE-ANR-2	ANR with invertible FET and colored noise model
IS	Itakura Saito
ITU	International Telecommunication Union
KLT	Karhunen-Loève transform
LPC	Linear prediction coefficient (or coding)
LSA	Log spectral amplitude
MBOM	Masking-based β -order MMSE
MBSD	Modified Bark spectral distortion
MSKF	Masking-based subband Kalman filtering
MMSE	Minimum mean-square error
PDF	Probability density function
PESQ	Perceptual evaluation of speech quality
PSD	Power spectral density
PWH-ANR	ANR with pre-whitening and invertible FET

QMF	Quadrature mirror filter
SANR	Subspace-based audible noise reduction
SDC	Spectral domain constraint
Seg.SNR	Segmental SNR
SNR	Signal-to-noise ratio
SPL	Sound pressure level
STSA	Short time spectral amplitude
SVD	Singular value decomposition
TDC	Time domain constraint

Symbols

α_1	Over subtraction factor
α_2	Spectral flooring factor
α	A factor for the permissible residual noise in TDC subspace enhancement technique
α_k	Phase of k th spectral component of a speech signal
α_p	A factor for the permissible residual noise in the SDC subspace enhancement technique
α_w	Weighting constant used for modified relative threshold offset
$\alpha_{t_1}, \alpha_{t_2}$	Constants which regulate the relation between noise PSD and masking threshold
β	Order of signal spectral amplitude using MMSE criterion
γ_k	<i>a posteriori</i> SNR
$\Gamma()$	Gamma function
δ	Nonnegative constant for μ value adjustment
$\tilde{\delta}$	Nonnegative constant for modified μ value adjustment
Δ_T	Time interval
ϵ_s	$\triangleq \text{tr}\left\{\text{E}\{d_s d_s^T\}\right\}$
$\varepsilon(k)$	Residual audible eigen-noise

$\tilde{\varepsilon}(k)$	Modified residual audible eigen-noise
ε	A very small positive constant
$\eta_n(k)$	Variance of k th spectral components of noise
$\eta_s(k)$	Variance of k th spectral components of a speech signal
$\eta(k)$	$\triangleq [\frac{1}{\eta_s(k)} + \frac{1}{\eta_n(k)}]^{-1}$
$\vartheta_{s,k}$	k th audible eigenvalue of a speech signal
$\vartheta_{\hat{s},k}$	k th audible eigenvalue of an estimated speech signal
$\iota(k)$	Constraint constant used in subspace speech enhancement
κ	Polyphase filter
λ_s	Eigenvalue of the autocorrelation of a speech signal
λ_t	Masking-eigenvalue
λ_p	p th eigenvalue
Λ_μ	Diagonal matrix of Lagrange multipliers
Λ	Local SNR
Λ_s	Eigenvalue matrix of the autocorrelation of a speech signal
μ_k	k th Lagrange multiplier
$\tilde{\mu}_k$	k th Lagrange multiplier for the case of colored noise
σ_w^2	Variance of $w(n)$
$\tilde{\sigma}_u^2$	Modified variance of $u(n)$
$\tilde{\sigma}_v^2$	Modified variance of $v(n)$
$\sigma_n^2(k)$	k th diagonal element of $U^T R_n U$
$\varsigma(l, k)$	$\triangleq \max[Z_k(l) - \sqrt{\eta_n(l, k)}, -\varepsilon]$
ϑ	Phase of k th spectral component of a noisy speech signal
Θ_f	Perceptual factor of the human auditory system in the frequency domain
ξ_k	<i>a priori</i> SNR
Ξ	Frame SNR
ϖ	Frame advance size

$\rho_1(k)$	$\triangleq \frac{T_f(l,k)-T_{min}(l)}{T_{max}(l)-T_{min}(l)}(\rho_{1max} - \rho_{1min}) + \rho_{1min}$
$\rho_2(k)$	$\triangleq \frac{T_f(l,k)-T_{min}(l)}{T_{max}(l)-T_{min}(l)}(\rho_{2max} - \rho_{2min}) + \rho_{2min}$
ρ_{1min}	Minimum of ρ_1
ρ_{1max}	Maximum of ρ_1
ρ_{2min}	Minimum of ρ_2
ρ_{2max}	Maximum of ρ_2
$\varrho_k(l)$	$\triangleq p(\mathbb{H}_0(k, l))$
τ_i	Polynomial coefficients
v_k	$\triangleq \frac{\xi_k}{1+\xi_k}\gamma_k$
Υ	$\triangleq \frac{1}{\gamma_k}$
ϕ_k	Phase of k th spectral component of a noisy speech signal
$\Phi(O_h)$	High-order polynomial terms
$\Phi(\beta_m, v_k)$	A simplified part of gain function of β -order MMSE estimator
$\psi(p)$	Power spectral density
$\tilde{\psi}_n$	Modified power spectral density of a noise signal based on masking threshold
ψ_s	Power spectral density of a speech signal
$\check{\psi}_s$	Power spectral density of the estimated speech signal at previous iteration
Ψ	Masking threshold
$\bar{\Psi}$	Normalized Masking threshold
$\Psi_{min}(l)$	Minimum of noise masking threshold $\Psi_f(l, k)$ at current frame l
$\Psi_{max}(l)$	Maximum of noise masking threshold $\Psi_f(l, k)$ at current frame l
Ψ_f	Normalized masking threshold in the frequency domain
$\Psi_s(j)$	Masking threshold corresponding to j th frequency bin of a subband
$\Omega_k(p)$	Coefficient of inverse frequency eigendomain transformation
a_i	i th Linear prediction coefficient
a_k	Variable to express A_k in integral equation
$\mathbf{a}$	Vector of linear prediction coefficients for a speech signal

A_k	Amplitude of the k th speech spectral component
$\mathbb{A}_m$	Arithmetic mean
$\mathbf{b}$	Vector of linear prediction coefficients of a colored noise signal
B_{bark}	The number of critical bands
$\mathbb{B}$	The number of subbands
C_{bark}	Critical Bark band number
C	Known measurement matrix
c_{ij}	Polynomial coefficients
d	Total distortion
$d_s(j)$	Speech distortion
$d_n(j)$	Residual noise
$\text{diag}()$	Diagonal matrix operator
$D(j_b, n)$	Estimate of noise with the smoothed periodogram
e	Base of natural logarithm
$\mathbf{E}$	Expectation operator
$\bar{f}_s$	Frequency bin number in a subband
$\bar{f}$	Frequency bin number in full-band
$\mathbb{F}$	Function operator
$\mathbf{F}$	Transition matrix
$\mathbf{F}_v$	Transition matrix
$\tilde{\mathbf{F}}$	Transition matrix
$\mathbf{g}$	Process noise coefficient vector
$\mathbf{g}_v$	Process noise coefficient vector
$\bar{\mathbf{g}}$	Process noise coefficient matrix
$g_k(n)$	k -th synthesis filter
g_k	k -th diagonal element of G
G_1	Full rank $M \times M$ matrix

$\mathbf{G}(n)$	Kalman gain at time n
$G_{\beta}(\xi_k, \gamma_k)$	Suppression gain function
$G_M(\xi_k, \gamma_k, \beta)$	Suppression gain function
$\mathbb{G}_m$	Geometric mean
G	$\triangleq \begin{bmatrix} G_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$
$\tilde{h}$	Prototype filter
$h_k(n)$	k -th analysis filter used in subband decomposition
H	Linear filter of subspace speech enhancement
$\mathbb{H}_0$	Speech-absence in a binary hypothesis model
$\mathbb{H}_1$	Speech-presence in a binary hypothesis model
i	i -th polyphase filter used in subband decomposition
$I_0()$	Modified Bessel function of zero order
$\mathbb{J}$	Cost function
K	The number of frequency bins
$\mathbf{K}(n, n-1)$	Autocorrelation matrix of predicted state-error
l	Frame number
$\mathbb{L}$	Up/down sampling rate
ℓ	ℓ th subband
$\max$	Maximum operator
$\min$	Minimum operator
M	Rank of autocorrelation matrix of a speech signal
$M(C_{bark})$	The number of frequency bins in critical Bark band C_{bark}
$\mathbb{M}(:,;)$	Confluent hypergeometric function
$n(t)$	Noise at time t
N_k	k th spectral component of a noise process
N_{sub}	The number of frequency bins in a subband

N_{full}	The number of frequency bins in full-band
$O_f(\lambda)$	Fixed relative threshold offset
$O_v(\lambda)$	Variable relative threshold offset
$O_m(k)$	Modified relative threshold offset
$p(X_k a_k, \alpha_k)$	Conditional probability density function of X_k given a_k and α_k
$p(a_k, \alpha_k)$	Joint PDF of the speech spectral amplitude and phase
$p(\mathbb{H}_0(k, l))$	Speech-absence probability
$p(\mathbb{H}_1(k, l))$	Speech-presence probability
p	Order of a signal autoregressive model
P	Vector dimension
q	Order of a colored noise autoregressive model
Q	Sparse matrix used in Kalman algorithm
Re	Real-part operator
$r(p)$	Autocorrelation function with lag p
$\hat{R}_{ss,n}$	Estimate of the autocorrelation function of $s(n)$
R_s	Autocorrelation matrix of a speech signal
R_{prew_x}	Autocorrelation function of the pre-whitened noisy signal
$R_{\hat{s}}$	Autocorrelation matrix of the estimated speech signal
$\tilde{R}_{N,n}(i)$	Modified autocorrelation of noise process with lag i
$s(t)$	Speech signal at time t
$\mathbf{s}(n)$	Vector of a speech signal at time n
S_k	k th spectral component of a speech signal
$\hat{\mathbf{s}}(n \mathbf{x}_n)$	Filtered estimate of the state vector $\mathbf{s}(n)$ at time n
$\hat{\mathbf{s}}(n \mathbf{x}_{n-1})$	One-step predicted estimate of the state vector $\mathbf{s}(n)$ at time n
S_{FM}	Spectral flatness measure
S_{func}	Spreading function of masking properties
$\check{S}(j_b, n)$	Fourier Transform of an estimated speech signal at previous iteration

t	time t
t_{one}	Tonality coefficient
$u(n)$	Excitation signal in colored noise model
$\mathbf{u}_p$	Eigenvector corresponding to the p th eigenvalue
U	Orthonormal eigenvector matrix of a speech signal
$\mathbf{v}(n)$	Vector of white Gaussian noise at time n
$V_p(k)$	k th value of Fourier transform of p th eigenvector
$w(n)$	Excitation signal in an AR speech model
$\mathbf{w}(l)$	Vector of white noise used in a subspace approach
w	Time-frequency window used in β -Order MMSE
$x(t)$	Noisy speech at time t
$\mathbf{x}(n)$	Vector of the observed noisy speech at time n
X_k	k th spectral component of an observed signal
Z_k	Amplitude of the k th noisy speech spectral component

List of Figures

2.1	Block diagram of speech enhancement system.	11
2.2	Block diagram of speech signal propagation.	11
2.3	A typical speech production model.	13
2.4	A simplified speech production model.	14
2.5	Block diagram of transform-domain based speech enhancement.	14
2.6	Block diagram of masking-based speech enhancement.	19
2.7	Fixed relative threshold offset.	22
2.8	Masking threshold computation.	24
2.9	An example of segmental SNR variation with time; for noisy speech with input average segmental SNR = -6.22 dB	27
2.10	An example of IS computation for a noisy speech with average IS = 3.13.	28
3.1	(a) Gain versus instantaneous SNR ($\gamma_k - 1$) in comparison with the Wiener gain (bold-dashed) for $\beta = 0.001$ (dotted), $\beta = 1.0$ (dotdash), $\beta = 2.0$ (dashed) and $\beta = 4.0$ (solid), and a priori SNR $\xi_k = -15, -5, 15$ dB. (b) Gain versus instantaneous SNR ($\gamma_k - 1$), for different β values and a priori SNR $\xi_k = -5$ dB.	37
3.2	(a) Gain versus β value for <i>a priori</i> SNR $\xi_k = -20, -10, 0, 10, 20$ dB and instantaneous SNR ($\gamma_k - 1$) = 0 dB. (b) Gain versus β value for instantaneous SNR ($\gamma_k - 1$) = -20, -10, 0, 10, 20 dB and <i>a priori</i> SNR $\xi_k = 0$ dB.	39
3.3	Gain curves for the cases of instantaneous SNR ($\gamma_k - 1$) equals to ξ_k for the Wiener gain, MMSE with $\beta = 0.001$ and $\beta = 4.0$	40

3.4	Segmental SNR comparison for the estimated speech signals obtained by the MMSE estimators with different β values and input noisy speech with seg.SNR (*) -10 dB, ($\diamond$) -5 dB, (x) 0 dB, (o) 5 dB, and (+) 10 dB.	43
3.5	PESQ comparison. Solid line for the estimated speech obtained by the MMSE estimators with different β values, and dashed line for input noisy speech with seg.SNR (*) -10 dB, ($\diamond$) -5 dB, (x) 0 dB, (o) 5 dB, and (+) 10 dB.	44
3.6	The adaptation of the values of β by using our proposed masking-based adaptation (a) The values of β adapted by using Eq. (3.44) at frequency 1000 Hz and 2000 Hz, for a speech utterance contaminated by F-16 noise with input segmental SNR of -10 dB; (b) The short-term spectral amplitude of the clean speech utterance at frequency 1000 Hz and 2000 Hz; (c) The theoretic frame SNR curve by using Eq. (3.18) and the estimated frame SNR by using Eq. (3.19) and the segmental SNR of -10 dB for the input noisy speech; (d) The waveform of the clean speech utterance.	55
3.7	The structure of the proposed masking-based adaptive β -order MMSE method.	56
3.8	Variations of <i>a priori</i> SNR ξ_k and instantaneous SNR ($\gamma_k - 1$) versus frame number. (k corresponds to 1 kHz.)	57
3.9	Speech spectrograms (male - "The fifth jar contains big, juicy peaches", 8 kHz sampling rate): (a) Clean Speech; (b) Noisy speech (F16 noise); (c) Wiener filtering method; (d) Over-subtraction masking method; (e) E-M LSA; (f) E-M MMSE; (g) Adaptive β -order MMSE; (h) Masking-based adaptive β -order MMSE.	60
3.10	Speech spectrograms (female - "The industry bargains as a unit with a single union", 8kHz sampling rate): (a) Clean speech; (b) Noisy speech (F16 noise); (c) E-M LSA; (d) OM-LSA; (e) Adaptive β -order MMSE; (f) Time-frequency adaptive β -order MMSE without speech-presence uncertainty; (g) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (h) Masking-based adaptive β -order MMSE.	61

3.11	Improvement in seg.SNR for white noise. The tested methods are as following: (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\square$) Optimal modified LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (Δ) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (*) Proposed masking-based adaptive β -order MMSE.	62
3.12	Improvement in seg.SNR for F16 cockpit noise. The tested methods are as following: (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\square$) Optimal modified LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (Δ) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (*) Proposed masking-based adaptive β -order MMSE.	63
3.13	PESQ measure for white cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.	64
3.14	PESQ measure for F16 cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.	65
3.15	Itakura-Saito distortion measure for white noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.	66
3.16	Itakura-Saito distortion measure for F16 cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.	67
4.1	(a) The frequency response of the prototype filter for the 16-band configuration; (b) The frequency response of the 16-band analysis filters; (c) The frequency response of the 16-band synthesis filters; (d) The frequency response of the overall 16-band system.	77
4.2	The flow chart for speech AR parameters estimation at the I -th iteration in the k -th subband.	80
4.3	An example of the power spectral magnitude of a speech signal (32 ms), and its spectral envelopes obtained by using AR parameters with $p=5$ and $p=16$ respectively.	80

4.4	The flow chart of the modified noise variance (or AR parameters for colored noise) estimation at the I -th iteration in the k -th subband. (a) white noise model; (b) colored noise model.	83
4.5	Structure of the proposed masking-based subband Kalman filtering scheme ('pre-IT' means the previous iteration, and IT denotes the number of iterations), where the double-lines represent the path of feedback from the previous iteration to the current iteration.	84
4.6	Vowel 'i' (8kHz sampling rate), with the input noisy SNR = 3.5 dB. (a) waveform of the vowel 'i', (b) ideal case AR spectral envelop, (c) zeroth iteration AR spectral envelop, (d) first iteration AR spectral envelop, (e) 3rd iteration AR spectral envelop, (f) 5th iteration AR spectral envelop.	85
4.7	consonant 's' (8kHz sampling rate), with the input noisy SNR = -10 dB. (a) waveform of the consonant 's', (b) ideal case AR spectral envelope, (c) zeroth iteration AR spectral envelop, (d) first iteration AR spectral envelop, (e) 3rd iteration AR spectral envelop, (f) 5th iteration AR spectral envelop.	86
4.8	Speech spectrograms: (a) Clean speech; (b) Noisy speech with input segmental SNR = 3.5 dB (white Gaussian noise); (c) Spectral Subtraction; (d) Over-subtraction masking method; (e) E-M MMSE method; (f) Full-band Kalman filtering; (g) Subband Kalman filtering method; (h) Proposed masking-based subband Kalman method.	95
4.9	Speech spectrograms: (a) Clean speech; (b) Noisy speech with input segmental SNR = -10.0 dB (Volvo car interior noise); (c) Spectral Subtraction; (d) Over-subtraction masking method; (e) Full-band Kalman filtering; (f) Full-band Kalman post-masking; (g) Subband Kalman filtering method; (h) Proposed masking-based subband Kalman method.	96
5.1	(a) Waveform of a speech signal and (b) the masking threshold.	104
5.2	Comparison of the original eigenvalues and the eigenvalues obtained by using our invertible FET and the conventional FET (a) the eigenvalues given the PSD and the eigenvector of the speech signal according to the FETs; (b) the masking-eigenvalues obtained by using the FETs as well as the eigenvalues obtained by the eigen-decomposition through the masking threshold.	104
5.3	The eigenvectors obtained by using: (o) speech signal and (*) the masking threshold.	105

5.4	Objective comparison for the case of white noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (-) Noisy speech (White noise), ($\square$) EV, ($\diamond$) JC-1, (o) JC-2, (*) CFE-ANR-1, (+) IFE-ANR-1.	118
5.5	Objective comparison for the case of F-16 (colored) noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (-) Noisy (F-16 cockpit noise), ($\square$) PWH-LE, (o) JC-2, (Δ) PWH-ANR, (*) CFE-ANR-2, (+) IFE-ANR-2.	119
5.6	Objective comparison for the case of Volvo (colored) noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (-) Noisy (F-16 cockpit noise), ($\square$) PWH-LE, (o) JC-2, (Δ) PWH-ANR, (*) CFE-ANR-2, (+) IFE-ANR-2.	120
5.7	Speech spectrogram of a clean speech.	121
5.8	Speech spectrograms of noisy and enhanced speech: (a) Noisy speech (White noise), (b) EV, (c) JC-1, (d) IFE-ANR-1.	122
5.9	Speech spectrograms of noisy and enhanced speech: (a) Noisy (F-16 cockpit noise), (b) JC-2, (c) PWH-ANR, (d) IFE-ANR-2.	123
6.1	Segmental SNRs of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and ($\diamond$) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.	134
6.2	IS distortion measurements of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and ($\diamond$) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.	135
6.3	PESQ scores of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and ($\diamond$) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.	136
6.4	MBSD measurements of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and ($\diamond$) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.	137

6.5	Spectrograms of (a) clean speech (female), (b) noisy speech contaminated by white noise, processed speech by using (c) $M\beta OM$, (d) MSKF, and (e) SANR.	138
6.6	Spectrograms of (a) clean speech (male), (b) noisy speech contaminated by colored noise (F-16 flight cockpit noise), processed speech by using (c) $M\beta OM$, (d) MSKF, and (e) SANR.	139
7.1	Power spectrum measure in a car	144
7.2	Curve of $\Delta(f)$ function	145
7.3	Speech waveforms and spectrograms: (a) Clean speech (b) Noisy speech (90 km/h car noise with window open) (c) LSA estimated speech signal (d) Modified $M\beta OM$ estimated speech signal. . .	147
7.4	Speech waveforms and spectrograms: (a) Clean speech (b) Noisy speech (115km/h car noise with window close) (c) LSA estimated speech signal (d) Modified $M\beta OM$ estimated speech signal. . .	148
7.5	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified $M\beta OM$ (star) with car noise at highway with window close at speech 70 km/h. . . .	149
7.6	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified $M\beta OM$ (star) with car noise at highway with window close at speech 90 km/h. . . .	150
7.7	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified $M\beta OM$ (star) with car noise at highway with window close at speech 115 km/h. . . .	151
7.8	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified $M\beta OM$ (star) with car noise at highway with window open at speech 70 km/h. . . .	152
7.9	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified $M\beta OM$ (star) with car noise at highway with window open at speech 90 km/h. . . .	153

7.10	Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window open at speech 110 km/h. . . .	154
7.11	The training phase and recognition phase of a typical speech recognition system.	155
7.12	Block diagram of performance evaluation for speech recognition	160
7.13	Speech recognition success rate in the presence of white noise.	161
7.14	Speech recognition success rate in the presence of colored (babble) noise.	161

List of Tables

2.1	MOS and corresponding speech quality	26
3.1	MOS values for the following methods. 1) WF: Wiener filtering; 2) LSA: E-M Log spectral amplitude; 3) MMSE: E-M MMSE; 4) OM: Over-subtraction based on masking properties; 5) β : Proposed adaptive β -order MMSE; 6) U- β : Proposed time-frequency adaptive β -order MMSE with speech presence uncertainty; 7) M- β : Proposed masking-based adaptive β -order MMSE. The input seg.SNR is -5 dB with MOS = 1.85.	68
4.1	IMPROVEMENT IN SEGMENTAL SNR OF ENHANCED SPEECH, FOR COMPARISON OF DIFFERENT BAND NUMBER, WITH NOISY INPUT SEGMENTAL SNR = -5 dB	87
4.2	IMPROVEMENT IN SEGMENTAL SNR OF ENHANCED SPEECH, FOR COMPARISON OF DIFFERENT AR ORDER (p, q) , WITH NOISY INPUT SEGMENTAL SNR = -5 dB, $\mathbb{B} = 16$	92
4.3	SEGMENTAL SNR IMPROVEMENT AND PESQ SCORE OF ENHANCED SPEECH FOR WHITE GAUSSIAN NOISE, FOR COMPARISON OF DIFFERENT EFFECTS OF FUNCTIONS, NOISY INPUT SEGMENTAL SNR = -5 dB AND PESQ = 1.63. FK: FULL-BAND KALMAN FILTERING METHOD($p=16$); SK: SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, p=1$); MSK: THE PROPOSED MASKING-BASED SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, p=1$).	92
4.4	SEGMENTAL SNR IMPROVEMENT AND PESQ SCORE OF ENHANCED SPEECH FOR COLORED (Volvo) NOISE, FOR COMPARISON OF DIFFERENT EFFECTS OF FUNCTIONS, NOISY INPUT SEGMENTAL SNR = -5 dB AND PESQ = 3.13. FK: FULL-BAND KALMAN FILTERING METHOD ($(p, q)=(16, 5)$); SK: SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, (p, q)=(1, 1)$); MSK: THE PROPOSED MASKING-BASED SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, (p, q)=(1, 1)$). . .	92

4.5	Comparison between full-band Kalman filtering ($p=16$) and the proposed method ($\mathbb{B}=16, p=1$), with White Gaussian Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB and 5 dB . . .	93
4.6	Comparison between full-band Kalman filtering ($(p,q)=(16,5)$) and proposed method ($\mathbb{B}=16, (p,q)=(1,1)$), with colored (Volvo) Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB and 5 dB	93
4.7	Segmental SNR improvement and PESQ of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($p=16, IT=4$); KM: Full-band Kalman post-masking ($(p=10), IT=4$); SK: Subband Kalman ($p=1, \mathbb{B}=16, IT=2$); and MSK: Our proposed Method ($p=1, \mathbb{B}=16, IT=2$) with White Gaussian Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB	94
4.8	Segmental SNR improvement and PESQ of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($(p,q)=(16,5), IT=4$); KM: Full-band Kalman post-masking ($(p=10), IT=4$); SK: Subband Kalman ($(p,q)=(1,1), \mathbb{B}=16, IT=2$); and MSK: Our proposed Method ($(p,q)=(1,1), \mathbb{B}=16, IT=2$) with Colored (Volvo) Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB	94
4.9	MOS score of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($(p,q)=(16,5), IT=4$); KM: Full-band Kalman post-masking ($(p=10), IT=4$); SK: Subband Kalman ($(p,q)=(1,1), \mathbb{B}=16, IT=2$); and MSK: Our proposed Method ($(p,q)=(1,1), \mathbb{B}=16, IT=2$) with Colored (Volvo) Noise, Noisy Input Segmental SNR: -5 dB with MOS = 2.18	94
5.1	Preference results in the case of white noise	121
5.2	Preference results in the case of F-16 noise	126
5.3	Preference results in the case of Volvo noise	127

Chapter 1

Introduction

1.1 Motivation

Speech enhancement has been a challenge for more than three decades. It mainly involves improving the performance of speech communication systems in noisy environments. Though research on speech enhancement has been going on for a long time, issues such as distortions to the original speech signals and residual noise created by the enhancement algorithms remain unsolved. A large number of research papers on different approaches and methods have managed to address these problems with varying degrees of success.

As the statistics of speech as well as noise signals are not explicitly available, and a truly reliable perceptually-based distortion measure is yet to be developed, model-based approaches continue to be extensively studied and applied to three main problems pertaining to speech enhancement. These problems are 1) speech estimation given a noisy speech, 2) speech coding when only a noisy speech is available, and 3) speech recognition given a noisy speech signal in man-machine communication.

For the purpose of enhancing speech with minimal distortion to the underlying speech signals and more perceptual noise reduction, masking properties of the human auditory system may be exploited. The performance of many speech enhancement schemes can be improved by exploiting masking properties. However, the problem of the optimal application of masking properties in speech enhancement is still open. It is well-known in the study of psychoacoustics that strong speech signals can mask weaker noise signals. This particular effect of the human auditory system

has been exploited in many audio and speech coding systems. Its use in speech enhancement systems is not wide spread and we attempt to re-examine it in our design of speech enhancement algorithms. Filtering a noisy speech signal usually causes distortions to the clean speech signal, though it may be successful in removing some of the noise. If we know that some parts of the noisy signal need not be filtered out because the noise components are masked by the clean speech signal, we can minimize the distortions introduced and produce a better enhanced speech signal. In this thesis, we focus on the effect of incorporation of masking properties into three speech enhancement methods.

Firstly, Gaussian-model Ephraim-Malah (E-M) MMSE and E-M log spectral amplitude (LSA) MMSE, that are based on the statistical independence assumption, are the most popular short-time spectral amplitude (STSA) algorithms. They outperform Wiener filtering and spectral subtraction, especially on elimination of musical noise. Both E-M MMSE and E-M LSA-MMSE estimators are considered as milestones of speech enhancement research. A number of research papers on speech enhancement methods are developed based on one of these two methods, especially E-M LSA-MMSE. In this thesis, a generalized Gaussian-model MMSE algorithm, i.e., β -order MMSE, is proposed. In fact, the E-M MMSE and E-M LSA-MMSE estimators are two special cases of β -order MMSE. Since the statistical independence assumption is not strictly satisfied for speech signal processing, weak speech spectral components cannot be effectively preserved. In this work, a novel adaptation of the value of β is proposed to overcome the weakness arising from the assumption. Through the adaptation of β value, the masking properties are incorporated into the β -order MMSE speech enhancement system.

Secondly, though Kalman filtering has been shown to be effective in speech enhancement, the potential benefits of incorporating masking properties into Kalman filtering for speech enhancement are still not sufficiently realized. The major stumbling block in the application of masking properties in this case is that Kalman filtering with the well-known speech production model, i.e., the autoregressive (AR) model, is realized in the time domain while the masking threshold is measured in the frequency domain. In this work, we propose a subband Kalman

filtering scheme to solve this problem. In addition, subband Kalman filtering has been shown to be better than fullband Kalman filtering in speech enhancement application in terms of many objective measurements and computational complexity.

Thirdly, subspace signal processing has been reported to have many advantages in signal processing. Although the incorporation of masking properties into the eigendomain has been applied through frequency eigendomain transformation (FET), the conventional FET is an approximate solution. In this work, an exact solution is proposed to replace the conventional one. As the audible and inaudible components are not differentiated in conventional masking-based subspace speech enhancement, an audible noise reduction subspace-based speech enhancement scheme is proposed.

From the practical and commercial point of view, the market potential of single channel speech enhancement could be enormous. For example, the speech quality in a moving vehicle could be severely degraded when a hands-free telephone is used due to the presence of unwanted noise. The degraded speech quality will cause digital cellular phone vocoders to malfunction, and produce distortion in the decoded speech. In addition, the implementation cost of single-channel speech enhancement is likely to decline with time and it is more compact as compared to multi-channel enhancement systems. Hence it will certainly benefit most users. On the other hand, robustness of a speech recognition system working in a noisy environment is an open challenge in research. To conclude this work, the effectiveness of the proposed speech enhancement methods are investigated for voice communication in a moving car environment as well as a particular speech recognition system.

1.2 Scope of the Thesis

In this thesis, we study single-channel speech enhancement methods based on masking properties. As shown by a large number of papers in the literature, the issues associated with enhancement using one microphone are theoretically challenging and practically important. With one microphone, the information available is much less than the case of two or more microphones. Consequently, noise suppression with one microphone often has to be effected through estima-

tion of noise and signal statistics, modelling of signal and noise, determination of model and statistical parameters, etc., which can be theoretically very involved.

The main objective of speech enhancement is to reduce the corrupting noise components of a noisy speech signal while preserving the original clean speech quality as much as possible. In this thesis, the development of various speech enhancement methods based on the properties of the human auditory system is focused, in order to improve the perceptual quality and intelligibility of given degraded speech signals, and also to increase the robustness of a speech recognition system to input noisy speech signals.

The thesis is organized as follows: in Chapter 2, key methods of speech enhancement as well as the masking properties of the human auditory system are reviewed. In addition, a modified relative threshold offset for noise masking threshold computation is described. In Chapter 3, the β -order MMSE estimation is introduced in detail for speech enhancement, where an adaptive β -order MMSE method with speech-presence uncertainty and a masking-based β -order MMSE method are further proposed. In Chapter 4, a masking-based Kalman filtering method is proposed, where masking properties are integrated into subband Kalman filtering scheme. Various investigations on subband Kalman filtering parameter setting through computer simulations and performance evaluation are provided. We further propose a subspace-based audible noise reduction method in Chapter 5. We introduce a novel invertible frequency eigendomain transformation, which is used to achieve audible noise reduction in the eigendomain, and to obtain a solution to the colored noise case. We then describe the implementation of the proposed enhancement method. In Chapter 6, the performance comparison of some of the proposed methods is given in terms of various objective measurements. In Chapter 7, two applications are studied using our proposed methods. One is a speech enhancement application for a voice communication system in a moving car environment; the other is speech enhancement for speech recognition. Evaluation results for both are provided and discussed. In Chapter 8, we conclude the thesis and propose recommendations for future research.

1.3 Contributions

The major contributions of this dissertation are summarized as follows:

1. Firstly, a novel β -order MMSE estimation is proposed. Following the analysis of the characteristics of β -order MMSE for speech enhancement, the effect of different fixed- β order MMSE is investigated. Moreover, an adaptive β -order MMSE, a time-frequency adaptive β -order MMSE method with speech-presence uncertainty as well as a masking-based β -order MMSE method are proposed. In addition, a simplification approach to reduce the computational complexity of β -order MMSE is elaborated.
2. A novel masking-based subband Kalman filtering scheme is proposed. The different functions including Kalman filtering, subband decomposition and perceptual effect based on masking threshold are studied. The optimal configuration for the proposed masking-based Kalman filtering method is investigated. The effectiveness of the proposed method is also evaluated.
3. A novel subspace-based audible noise reduction method and an invertible frequency eigen-domain transformation are proposed. The implementation of the proposed signal subspace scheme is outlined and extended to the case of colored noise.
4. A modified relative threshold offset approximation for masking threshold estimation is proposed.
5. The masking-based β -order MMSE method applied to a moving car environment and the proposed masking-based Kalman filtering applied to a speech recognition system are also investigated.

Chapter 2

Background of Masking-Based Speech Enhancement System

2.1 A Brief History of Speech Enhancement

Speech enhancement mainly deals with the improvement of speech quality, intelligibility and robustness for further analysis or synthesis. Since the key concern is the intelligibility of the enhanced speech, it would generally be acceptable to sacrifice the quality of speech if its intelligibility could be improved. Although the speech signal can be degraded by multiplicative noise or convolutional noise, these cases not only seldom occur in practice but also can be converted to additive noise using a homomorphic transformation. In this section, we review some of the well known enhancement methods for processing speech signals degraded by additive noise, where the only signal available is degraded speech and the noise is independent of the original speech. The problem of speech enhancement that involves mainly noise reduction remains a key issue in communication. A great number of techniques have already been proposed and examined.

In the earlier years from 1940s to 1970s, speech enhancement began to attract the interest of many researchers. During that period, researchers only focused on the suppression of noise for direct listening purposes. In 1947, Licklider and Pollack [63] reported their investigation on the effect of clipping in single microphone speech enhancement. As one knows, clipping can cause the generation of harmonics and in turn could improve speech intelligibility despite also introducing some undesirable distortions. The problem of speech enhancement has received considerable attention in the literature since the mid-1970s. In 1974 Weiss *et al.* [73] developed an earlier

version of spectral enhancement algorithm for the Italian Airforce, and the underlying algorithm was later identified to be a variant of spectral subtraction, which became very popular after the work of Boll in 1979 [98]. A pioneering contribution to the theory and applications of speech enhancement was made by Lim and Oppenheim, in their landmark paper [48] in 1979, wherein they formulated the speech enhancement problem, classified the different speech enhancement systems known at that time according to the principles they were based upon, and evaluated and compared these systems. Musical tone noise in spectral subtraction motivated appreciable research on exploring alternative methods. In 1984, Ephraim and Malah proposed an effective minimum mean-square error (MMSE) of short-time spectral amplitude (STSA) estimator [115], which not only produced more noise reduction but also reduced the level of tonal noise as compared to spectral subtraction.

In a conventional speech production model, speech is generated by exciting a quasi-stationary linear system with a pulse train for voiced speech and noise for unvoiced speech. Based on this model, a possible approach to speech enhancement is to estimate the parameters of the model rather than the speech signal itself and to use them to synthesize the speech signal, i.e., to enhance speech through the use of an analysis-synthesis system. In 1973, N.J. Miller used this concept to remove the orchestral accompaniment from early recordings of Enrico Caruso [76]. In 1978, Lim and Oppenheim presented several algorithms that estimate the all-pole vocal tract model parameters from degraded speech based on the maximum likelihood approach [46]. In 1987, Paliwal and Basu presented a Kalman filtering method for speech enhancement based on the time-varying autoregressive model (AR) [59]. Unlike spectral-domain enhancement, AR Kalman filtering is based on the production model of speech signal. It introduces additional human speech features to achieve some effects which cannot be substituted by the spectral-domain methods. In this approach, speech is modelled by the response of a linear system representing the vocal tract, which is driven by an excitation signal.

Two years later (or in 1989), Ephraim, Malah and Juang reported the use of the Hidden Markov Model (HMM) technique in speech enhancement [116]. Since many speech recognition

systems employ HMM techniques, the enhanced outcome could be considered for integration with those HMM speech recognition systems for applications under noisy environments. In 1998, Hossein Sameti *et al.* demonstrated that the use of general statistical characteristics of speech, as partially captured by the HMM trained from a large corpus of clean speech data, is beneficial in improving the performance of speech enhancement systems.

In 1993, Ephraim and Van Trees came up with a signal subspace approach for speech enhancement [114]. The underlying principle is to first decompose noisy speech into speech-plus-noise subspace and noise subspace. Enhancement is then carried out via suitable estimation techniques over the speech-plus-noise subspace. It results in enhanced speech that is free of musical tones. Recently several subspace-based speech enhancement methods have been proposed, following the pioneering work of Ephraim and Van Trees, which dealt primarily with white noise and suggested using pre-whitening to handle colored noise. In 2001, Rezayee and Gazor extended Ephraim and Van Trees' time domain constrained method to deal with colored noise by approximating the covariance matrix of the KLT-transformed noise vectors with a diagonal matrix [1]. Their approximation, however, led to a sub-optimal estimator.

In 1933, Fletcher and Munson established the curve of equal loudness, which shows a distinct frequency response of the human hearing system [28]. The human ear is most sensitive in the frequency range between 2 and 4 kHz, and becomes less sensitive at other frequencies. In 1956, Stevens introduced a method for calculating a loudness index [103]. In 1961, Zwicker developed a hearing based method for noise measurement which calculates critical band loudness and total loudness [17]. He showed that the inner ear resolves sound with the help of mechanical filtering, with a bandwidth which he called critical bandwidth. This bandwidth corresponds to 100 Hz absolute bandwidth for frequencies below 500 Hz, and to approximately 20 % relative bandwidth for frequencies above 500 Hz. An important consideration in speech enhancement systems arises from the fact that the criteria for enhancement ultimately relate to an evaluation by a human listener. In different contexts, the criteria for evaluation may differ depending on whether quality, intelligibility, or some other attributes are more important. Thus speech enhancement must

inevitably take into account aspects of human perception. Human auditory modeling has been widely used in acoustic signal processing, especially in audio [84] and speech coding [4]. This model is based on a central analysis mechanism of the inner ear, which reveals the essential principle of masking. Masking and bandpass filtering represent two simple ways in which perceptual aspects of the human auditory system can be exploited in speech enhancement. In 1968, Thomas and Niederjohn [34] applied high-pass filtering to weaken the first formant followed by infinite amplitude clipping. The high-pass filtering increases the amplitude of higher formants relative to the first formant for vowels. It is motivated by an observation that the second formant gives more significant effect for intelligibility than the first formant [35]. In addition, they observed that infinite amplitude clipping can increase the amplitude of the weak speech components such as consonants such that they are less susceptible to be masked by noise. Since weak speech spectral components such as consonants often convey significant information that affects intelligibility [26], some degree of preservation of these components during the enhancement process is becoming worthy of investigation. In 1993, D. Tsoukalas et al. incorporated the psychoacoustic criteria into speech enhancement techniques [14]. The masking properties are modelled by the noise masking threshold. Masking is present because the auditory system is incapable of distinguishing two signals which are close to one another in spectral spacing. Masking effects occur not only when sounds are presented simultaneously, but also when they are not.

2.2 Approaches of Speech Enhancement

In speech communication, information can usually be exchanged easily and accurately when the speaker and the listener are in a quiet environment. However, the presence of background interference always leads to the degradation of the quality and intelligibility of the desired speech signal. The interference can involve the signal conversion error from microphones, background noise, channel transmission noise, low fidelity of loudspeakers and headphones, and competing speech in a multi-speaker environment. There are various speech enhancement systems for a great variety of applications, which may be partitioned as single channel and multichannel. For single channel application, the characterization of noise statistics must be obtained during periods of

silence and based on the stationarity assumption of the noise. In many situations, such as mobile communication, voice telephony and preprocessing for speech recognition, only single channel noisy speech is available. A multichannel system is different from the single channel case in that it may contain the desired speech with additive noise in a primary channel, and the noise signal correlated to the noise of the primary channel in the other channel(s). Adaptive filtering approaches like acoustic echo cancellation (AEC) [42], acoustic noise reduction (ANR) [74] are well known in this category. Due to speech leakage to other channel(s) from the primary channel, speech enhancement sometimes need to address the issue of cross-talk or make use of an array beamforming solution.

In this thesis, single channel speech enhancement is studied. So far, the most popular methods of single channel speech enhancement can be broadly classified as follows:

1. Spectral suppression [36, 47, 48, 98, 115];
2. Speech production model based Kalman filtering [53, 59, 99, 110, 122];
3. Perceptual properties based speech enhancement [15, 20, 52, 121, 80];
4. Speech-presence uncertainty based speech enhancement [13, 37, 89];
5. Hidden Markov model (HMM) based speech enhancement [32, 116];
6. Enhancement using sinusoidal analysis and synthesis [43, 44, 69];
7. Signal subspace [21, 22, 70, 117, 120] and singular value decomposition (SVD) [3, 58, 100] based speech enhancement;
8. Wavelet speech enhancement [38, 49, 67].

Speech enhancement system attempts to obtain a good estimate of the clean speech signal from its noisy version. The problem of speech enhancement can be formulated as follows. The given noisy signal is modeled in an additive form, i.e.,

$$x(t) = s(t) + n(t) \quad (2.1)$$

where $x(t)$, $s(t)$ and $n(t)$ represent the noisy speech, clean speech and additive noise at time t respectively. The objective is therefore to devise an enhancement system to obtain the estimate of the clean speech. A typical block diagram of speech enhancement is shown in Fig. 2.1.

Figure 2.1: Block diagram of speech enhancement system.

There are many ways to categorize speech enhancement methods. A broad grouping is concerned with the manner in which the speech is modeled. Some techniques are based on stochastic process models of speech and noise, while others are based on perceptual aspects of speech. Enhancement systems based on stochastic process models depend on their corresponding criteria, and systems based on the perceptual properties attempt to emphasize something important in human perception. The propagation of a speech signal can be viewed as starting from the physical speech production system of a speaker, then going through the sound propagation media where possible noise contaminates the speech signal and finally reaching the receiver, e.g., human listener(s) or a machine speech recognition system. Along with signal propagation, we may group the bases of speech enhancement techniques into three models: the speech production model, the signal characteristics model and the speech receiver model. Fig. 2.2 shows the three phases of speech signal propagation and their corresponding models.

Figure 2.2: Block diagram of speech signal propagation.

2.2.1 The Speech Production Model

Speech is the acoustic wave originating from the sub-glottal system that is composed of lungs, bronchi and trachea. The airflow originating from the lungs is perturbed by the constriction of the vocal tract. The vocal tract consists of the oral tract and the nasal tract. Sounds of a human speaker which have airflow through the nasal cavity are called nasal sounds. Sounds which have no nasal airflow are called oral sounds. The area of the oral tract is determined by the positions of the tongue, lips, jaw and velum. It varies from 0 to about 20 cm^2 . The nasal tract begins at the velum which is a soft palate at the back of the roof of the mouth and ends at the nostrils. When the velum is lowered to create a velic opening to the nasal cavity, the nasal sounds of speech are acoustically produced through nasal tract coupling the vocal tract. When the velum is raised to form a velic closure with the pharynx, oral sounds are produced. Speech sounds of a human speaker can be categorized into three classes according to the mode of excitation. The first class is the voiced sounds produced by forcing air through the glottis with the tension of the vocal cords adjusted. The sounds vibrate in a relaxation oscillation to produce quasi periodic pulses of air which excite the vocal tract. The second is fricative or unvoiced sounds. They are actually the turbulence airflow with a high enough velocity that is generated by forming a constriction somewhere along the vocal tract. Subsequently, a broad-spectrum noise source to excite the vocal tract is created. The third is plosive sounds which are resulted from making a complete closure, building up pressure behind the closure, and abruptly releasing it. The oral tract and nasal tract can be outlined as tubes of non-uniform cross-sectional area. As sounds propagate down these tubes, the frequency selectivity of the tube shapes the frequency spectra in the way similar to the resonance effects observed with organ pipes or wind instruments. The resonance frequencies of the vocal tract are referred to as formant frequencies. The formant frequencies are determined by the shape and dimensions of the vocal tract. Each shape represents a set of formant frequencies. Different sounds are generated by varying the shape of the vocal tract. Hence the spectral properties of the speech signal vary with time as the shape of the vocal tract changes.

In a typical speech production model, speech is described as the acoustic product of well-controlled movements of the speech apparatus. In the past three decades, the technique of linear prediction has been proven to be very significant in providing an efficient representation of speech signal for a variety of speech processing applications. It is well known that linear prediction models the vocal tract. The linear prediction coefficient (LPC) model is based on the fact that a signal conveying messages is never completely random. There is a correlation between successive samples. The LPC model uses this correlation to reduce the amount of data while keeping the information in the signal. Most LPC systems assume an all-pole model, which is statistically an autoregressive (AR) model of the vocal tract. Fig. 2.3 shows a typical speech production model based on LPC above.

Kalman filtering based speech enhancement has been proven to be a successful application of an AR model in various research papers [7, 53]. In this case, the AR model is often used to construct the transition matrix in the state space. Consequently, the speech production model can be simplified into Fig. 2.4. Many Kalman filtering speech enhancement approaches use a white noise model with time-varying variance as the speech excitation signal [78, 99, 110]. In [122], a voiced/unvoiced model is considered in order to estimate more accurately voiced speech with the Kalman filter approach. However, the computational complexity is obviously increased.

Figure 2.3: A typical speech production model.

Figure 2.4: A simplified speech production model.

2.2.2 The Signal Characteristics Model

Almost all speech enhancement approaches need to exploit the signal characteristics model to estimate the speech signal and to reduce noise. Signal subspace speech enhancement, spectral suppression and HMM speech enhancement mainly exploit the statistical characteristics of the speech and noise signals to achieve speech enhancement. The statistical characteristics model may be represented by different probability distribution functions of the speech and noise signals, the dissimilar stationarity of speech and noise processes, the time-varying statistics in a certain transform domain, etc.

Fig. 2.5 illustrates the procedure of the transform-domain approaches which use the statistical characteristics of the speech and noise signals.

Figure 2.5: Block diagram of transform-domain based speech enhancement.

Spectral Subtraction

It is well recognized that spectral suppression is a rather effective way to realize the enhancement of a noisy speech signal [6, 13, 36, 37, 39, 68, 89, 98, 115, 118]. Spectral subtraction is a typical method of spectral suppression, which models the short-term stationary characteristic of speech and long-term stationary characteristic of noise for each spectral amplitude. With additive noise

degradation, we have the estimate of the clean speech spectral amplitude as:

$$\hat{S}(k) = X(k) - \bar{N}(k), \quad k = 0, \dots, K - 1 \quad (2.2)$$

where K is the number of frequency bins. $\bar{N}(k)$ is the expectation of $N(k)$ which may be estimated during speech-absence periods. $\hat{S}(k)$ can further be written in the following form:

$$\hat{S}(k) = G(k)X(k), \quad k = 0, \dots, K - 1 \quad (2.3)$$

where $G(k) = 1 - \frac{\bar{N}(k)}{X(k)}$ is the suppression gain. The enhanced speech signal is obtained by inverse transforming the estimated speech spectrum which is composed of the short-term estimated speech spectral amplitude and the phase of the noisy speech spectrum, as it has been found that the human ear is not sensitive to the phase of speech spectrum. From (2.2), it can be noted that the estimated speech spectral amplitude is not guaranteed to be positive. Many solutions have been proposed to rectify this problem and they include half-wave rectification [98], full-wave rectification [33], and floor weighting [80]. However, forcing a negative spectral amplitude value to an artificial value results in musical tone artifacts. Spectral subtraction therefore remains a merely reference method for performance comparison for other more advanced methods.

Signal Subspace

Signal subspace based speech enhancement methods also exploit a similar stationarity model through the Karhunen-Love transform (KLT). This model exploits the relative long-term stationarity of noise eigenvalues and short-term stationarity of speech eigenvalues [1, 114, 117, 120, 121]. Its suppression gain works in the eigen-domain.

In [117], Ephraim and Van Trees considered the speech enhancement problem in terms of linear estimation of the clean signal from the corrupted speech in the KLT domain. Their main contribution was to separate the estimation error into speech distortion and residual noise, and optimize the tradeoff between the two errors through the time domain constraint and the spectral domain constraint methods.

A noisy speech sequence $\mathbf{x}$ is assumed to be the sum of zero mean clean speech sequence $\mathbf{s}$

and a white noise sequence $\mathbf{w}$ with variance σ_w^2 , i.e.,

$$\mathbf{x} = \mathbf{s} + \mathbf{w} \quad (2.4)$$

Let R_s denote the autocorrelation matrix of clean speech. The eigen-decomposition of R_s is given by

$$R_s = U\Lambda_s U^T \quad (2.5)$$

where $\Lambda_s = \text{diag}[\lambda_{s,0}, \dots, \lambda_{s,K-1}]$ and $U = [\mathbf{u}_0, \dots, \mathbf{u}_{K-1}]$ are the eigenvalue matrix and the orthonormal eigenvector matrix of R_s respectively, and K is the vector dimension.

Let $\hat{\mathbf{s}} = H\mathbf{x}$ be a linear estimator of clean speech vector $\mathbf{s}$, where H is a $K \times K$ matrix. The residual distortion after this estimation is obtained by [117]

$$\mathbf{d} = \hat{\mathbf{s}} - \mathbf{s} = (H - I)\mathbf{s} + H\mathbf{w} \triangleq \mathbf{d}_s + \mathbf{d}_w. \quad (2.6)$$

where $\mathbf{d}_s$ and $\mathbf{d}_w$ represent signal distortion and residual noise respectively.

- TIME DOMAIN CONSTRAINED ESTIMATOR

In [117], a time domain constrained (TDC) estimator is proposed to find an optimal H in order to minimize the signal distortion over all linear filters which result in the permissible residual level. The linear estimator with TDC is defined by

$$\begin{aligned} H_{TDC} &\triangleq \arg \left\{ \min_H \text{tr} \left\{ \mathbb{E} \{ d_s d_s^T \} \right\} \right\} \\ \text{Subject to : } &\frac{1}{P} \text{tr} \left\{ \mathbb{E} \{ d_n d_n^T \} \right\} \leq \alpha \sigma_w^2 \end{aligned} \quad (2.7)$$

where $0 < \alpha < 1$, and P is the vector dimension. With the solution of the gradient equation of the Lagrangian, the optimal H is given by [117]

$$\begin{aligned} H_{TDC} &= R_s (R_s + \mu \sigma_w^2)^{-1} \\ &= U \Lambda_s (\Lambda_s + \mu \sigma_w^2)^{-1} U^T \\ &= U G U^T \end{aligned} \quad (2.8)$$

where μ is the Lagrange multiplier, and $G \triangleq \Lambda_s (\Lambda_s + \mu \sigma_w^2)^{-1}$.

As mentioned in [117], there is strong empirical evidence to indicate that the covariance matrices of most speech vectors have some eigenvalues which are practically equal to zero. Thus, the eigenvalue matrix Λ_s has only M non-zeros, where $M \leq P$, i.e.,

$$\Lambda_s = \text{diag}[\lambda_{s,0}, \dots, \lambda_{s,M-1}, 0, \dots, 0]. \quad (2.9)$$

Hence, (2.8) can be re-written as

$$H_{TDC} = UGU^T = U \begin{bmatrix} G_\mu & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^T. \quad (2.10)$$

where G_μ has a rank M . Consequently, the enhanced speech is obtained by firstly decomposing the noisy signal into a signal-plus-noise subspace and a noise subspace with the use of the Karhunen-Loève transform (KLT), followed by modifying the KLT coefficients by a gain function, and finally by applying inverse KLT on the modified coefficients to obtain the enhanced speech.

• SPECTRAL DOMAIN CONSTRAINED ESTIMATOR

The spectral domain constrained (SDC) estimator is derived to minimize the signal distortion subject to the constraints on the spectrum of the residual noise [117]. In particular, the linear filter H is defined by

$$H_{SDC} \triangleq \arg \left\{ \min_H \text{tr} \left\{ \mathbb{E} \{ d_s d_s^T \} \right\} \right\} \quad (2.11)$$

Subject to:

$$\begin{aligned} \mathbb{E} \left\{ \left| \mathbf{u}_p^T d_n \right|^2 \right\} &\leq \alpha_p \sigma_w^2, \quad p = 0, \dots, M-1 \\ \mathbb{E} \left\{ \left| \mathbf{u}_p^T d_n \right|^2 \right\} &= 0, \quad p = M, \dots, P-1. \end{aligned} \quad (2.12)$$

A possible solution of SDC is that G is a diagonal matrix given by

$$G = \Lambda_s (\Lambda_s + \sigma_w^2 \Lambda_\mu)^{-1} \quad (2.13)$$

where $\Lambda_\mu = \text{diag}[\mu_0, \dots, \mu_{K-1}]$ is a diagonal matrix of Lagrange multipliers, and the k -th diagonal element of G is given by [117]

$$g_k = \begin{cases} \frac{\lambda_{s,k}}{\lambda_{s,k} + \mu_k \sigma_w^2}, & k = 0, \dots, M-1 \\ 0, & k = M, \dots, K-1. \end{cases} \quad (2.14)$$

In [117], an alternative choice for the gain function, q_k , is proposed to replace g_k and it is given by

$$q_k = e^{-\nu \sigma_w^2 / \lambda_{s,k}}, \quad k = 0, \dots, K-1 \quad (2.15)$$

where ν is a constant that affects the suppression level of noise and the distortion level of speech signal.

2.2.3 The Speech Receiver Model

The speech receiver model exploits the features of human listeners and speech recognition systems. In the latter case, the features used for speech recognition are enhanced for the purpose of raising the speech recognition success rate when in a noisy environment [5, 30, 56, 61, 108]. In the former case, speech enhancement exploits the masking properties of the human auditory system to improve the perceptual quality of speech for human listeners. The use of masking properties to improve the quality and intelligibility as well as speech recognition success rate are reported in many publications. In this thesis, we focus on the human listeners as the receivers.

A typical speech enhancement system [7, 10, 14, 15, 21, 22, 78, 79, 80] that exploits masking properties is described in Fig. 2.6. It is believed that the masking properties can be exploited to better balance the introduction of artifacts and the reduction of noise. The masking properties cannot be used directly to suppress noise. Fortunately, they can be exploited to modify the design and/or the parameters of many existing speech enhancement algorithms. A robust speech enhancement system is usually established by using the combination of the above three models. As a result, a receiver model can only be realized by incorporating at least either a speech production model or a signal characteristics model.

Subtractive-type Speech Enhancement Based on Masking Properties

In [80], a generalized spectral subtraction algorithm is modified by varying the subtraction parameters depending on the masking threshold. In particular, the generalized spectral subtraction

Figure 2.6: Block diagram of masking-based speech enhancement.

scheme is designed with the following gain function:

$$G(k) = \begin{cases} \left(1 - \alpha_1 \left[\frac{|N(k)|}{|X(k)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{if } \left[\frac{|N(k)|}{|X(k)|}\right]^{\gamma_1} < \frac{1}{\alpha_1 + \alpha_2} \\ \left(\alpha_2 \left[\frac{|N(k)|}{|X(k)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{otherwise.} \end{cases} \quad (2.16)$$

where k is the frequency bin number and α_1 is the over-subtraction factor since $\alpha_1 > 1$ [80]. This means that the short-term spectrum is suppressed more than what is necessary. α_2 is referred to as spectral flooring factor with its range of $[0 \ 1]$, and it determines the minimum value of the gain function. The exponential factors γ_1 and γ_2 affect the sharpness of the transition from $G(k) = 1$ to $G(k) = 0$. In [80], $\gamma_1 = \frac{1}{\gamma_2} = 2$ is recommended through several experiments, and the incorporation of the perceptual properties is done by adapting the subtraction parameters α_1 and α_2 to the masking threshold Ψ . The adaptation of the subtraction parameters is provided with the following relations:

$$\alpha_1(k) = \mathbb{F}_1[\alpha_{1,min}, \alpha_{1,max}, \Psi(k)] \quad (2.17)$$

$$\alpha_2(k) = \mathbb{F}_2[\alpha_{2,min}, \alpha_{2,max}, \Psi(k)] \quad (2.18)$$

where $\alpha_{1,min}$ and $\alpha_{1,max}$ denote the minimum and maximum values of the over-subtraction scheme respectively; $\alpha_{2,min}$ and $\alpha_{2,max}$ denote the minimum and maximum values of the spectral floor respectively; $\alpha_{1,min} = 1$, $\alpha_{1,max} = 6$, $\alpha_{2,min} = 0$ and $\alpha_{2,max} = 0.02$ are chosen to obtain a good tradeoff between speech distortion and residual noise [80]. $\mathbb{F}_1$ and $\mathbb{F}_2$ denote functions leading to a maximal residual noise reduction for minimal values of the masking threshold and a minimal reduction for the maximal values of the threshold, i.e., $\mathbb{F}_1 = \alpha_{1,max}$ when $\Psi(k) = \Psi(k)_{min}$, and $\mathbb{F}_1 = \alpha_{1,min}$ when $\Psi(k) = \Psi(k)_{max}$, where $\Psi(k)_{min}$ and $\Psi(k)_{max}$ are the minimum

and maximum values of the masking threshold of the current frame. The values of $\mathbb{F}_1$ between the two extreme values are then interpolated depending on $\Psi(k)$. A similar approach is applied to $\mathbb{F}_2$.

Consequently, the gain function $G(k)$ is adapted indirectly by the masking properties, and the estimate of the speech spectral amplitude is obtained by multiplying the gain function with the noisy speech spectral amplitude.

2.3 Masking Properties of Human Auditory System

2.3.1 Masking Threshold

The analysis bandwidth of the human auditory system has been measured by Zwicker in detail [18]. The human auditory model is based on the masking phenomenon. It is related to the concept of critical band analysis, which is the central analysis mechanism of the inner ear. The masking properties are modelled by the noise masking threshold. Masking is present because the auditory system is incapable of distinguishing two signals which are close to one another in terms of frequency spacing. Masking effects occur not only when sounds are presented simultaneously, but also in cases when they are not. There are two non-simultaneous masking effects, i.e., premasking and postmasking. Premasking takes place during the period of time before the masker is present, while postmasking follows simultaneous masking. Noise masking is a well-known psychoacoustic property of the human auditory system that has already been applied with success to speech and audio coding in order to partially or totally mask the distortion introduced in the coding process. In this thesis, only frequency domain masking, or simultaneous masking is considered, where a weak signal is made inaudible by a stronger signal occurring simultaneously. This phenomenon is modelled via a noise masking threshold, below which all components are inaudible [14, 93].

Sounds require a minimum pressure in order to be audible. Moreover, due to the filtering effect of the outer and middle ear, this minimum pressure, which gives rise to the threshold of hearing, varies considerably with frequency. This threshold of hearing (audibility) is referred to as the absolute threshold of hearing. It is unique from person to person and changes with age.

In signal processing, the absolute threshold is approximated by:

$$\Phi_A(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \quad (dB \text{ SPL}) \quad (2.19)$$

which is measured in dB sound pressure level (SPL), or dB relative to $20 \mu Pa$ [55]. f is the linear frequency in Hz. The Bark frequency scale can be approximated by the following expression:

$$C_{bark}(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left[\left(\frac{f}{7500}\right)^2\right] \quad (2.20)$$

where C_{bark} denotes critical Bark band.

The noise masking threshold can be obtained through modelling the frequency selectivity of the human auditory system and its masking properties. The noise masking threshold can be obtained approximately by critical band analysis, a spreading function, the relative threshold offset and the absolute auditory threshold. The procedure of computing a noise masking threshold, $\Psi(C_{bark})$ consists of the following steps [55]:

1. Critical band analysis of speech signal to obtain the spectral power of each critical band ¹.
2. Obtain the spread masking threshold by convolving the critical band spectrum with the spreading function, S_{func} , which is given as follows:

$$S_{func}(C_{bark}) = 15.81 + 7.5(C_{bark} + 0.474) - 17.5\sqrt{1 + (C_{bark} + 0.474)^2} \quad (2.21)$$

3. Calculation of a relative threshold offset depending on the tonality of the pre-estimated speech signal.
4. The critical band masking threshold is obtained by subtracting the spread masking threshold from the relative threshold offset.
5. Normalization and comparison with the absolute threshold of hearing.

2.3.2 Modified Relative Threshold Offset

It is observed that the performance of a speech enhancement algorithm suffers greatly when the masking threshold is not correctly estimated. The masking threshold is used to achieve spectral

¹As we only have the noisy speech signal, we have to pre-estimate the speech signal by using certain pre-enhancement method.

Figure 2.7: Fixed relative threshold offset.

shaping of the artifacts arising from the noise removal process such that they are inaudible to the human ear. Although noise can be removed from speech in its original form, it may reappear in the form of equally disturbing tonal artifacts when the masking threshold is not correctly estimated.

There are two ways to obtain the relative threshold. One is given by a simple estimation which is based on the fact that the speech signal has a tone-like nature in the lower critical bands and a noise-like nature in the higher bands [80]. It can be represented using a fixed relative threshold offset, $O_f(C_{bark})$, shown in Fig. 2.7 [80]. It is also observed that the fixed relative threshold offset, $O_f(C_{bark})$, is generally suitable for most speech segments. Another approach is based on an estimation of the tonality in a particular critical Bark band, and the corresponding relative threshold offset is then obtained. The noise masking threshold calculation considers noise-masking tone and tone-masking noise [93, 71]. The noise-masking tone is estimated as $(14.5 + C_{bark})$ dB below the spread critical spectrum. The tone-masking noise is estimated to be 5.5 dB below C_{bark} uniformly across the spread critical spectrum. In order to apply the properties of the noise-masking tone and the tone-masking noise, the spectral flatness measure (S_{FM}) [93]

is used to determine if the signal is close to noise or tone. S_{FM} is defined by the ratio of the geometric mean ($\mathbb{G}_m$) to the arithmetic mean ($\mathbb{A}_m$) of the power spectral density components ($\psi(j)$) in each critical band, i.e.,

$$S_{FM}(C_{bark}) = 10 \log_{10} \frac{G_m(C_{bark})}{A_m(C_{bark})}, \quad C_{bark} = 1, 2, \dots, B_{bark}. \quad (2.22)$$

$$\mathbb{G}_m(C_{bark}) = \prod_{j=1}^{M(C_{bark})} (\psi(j))^{\frac{1}{M(C_{bark})}}, \quad \mathbb{A}_m(C_{bark}) = \frac{\sum_{j=1}^{M(C_{bark})} \psi(j)}{M(C_{bark})}, \quad (2.23)$$

where B_{bark} is the number of critical bands and $M(C_{bark})$ denotes the number of frequency bins in each band, C_{bark} . A coefficient of tonality, $t_{one}(C_{bark})$, is defined as

$$t_{one}(C_{bark}) = \max\{\min[\frac{S_{FM}(C_{bark})}{S_{FMmax}}, 1], 0\}. \quad (2.24)$$

Here S_{FMmax} is set to -60 dB. S_{FM} of -60 dB represents a pure tone-like signal, and S_{FM} of 0 dB indicates a completely noise-like signal. The variable relative threshold offset, $O_v(C_{bark})$, in each critical band is given by

$$O_v(C_{bark}) = t_{one}(C_{bark})(14.5 + C_{bark}) + [1 - t_{one}(C_{bark})]5.5. \quad (2.25)$$

Since we only have the noisy speech signal, the preliminary estimate of the speech signal is usually not very accurate. Consequently, the estimation of the variable relative threshold offset is also affected. This is partially because the residual noise of the preliminary estimated speech may change the original tonality of the speech signal. As a result of the inaccurate relative threshold offset, the estimation error of the masking threshold increases. In [80], the fixed relative threshold offset is exploited and compensated with a modification by taking into account the tone-like nature of the musical residual noise for $C_{bark} > 15$. We propose here a modified offset by merging both the fixed and variable offsets to achieve a more effective application of the masking properties. The modified relative threshold offset, $O_m(k)$, is given by

$$O_m(C_{bark}) = \alpha_w O_v(C_{bark}) + (1 - \alpha_w) O_f(C_{bark}) \quad (2.26)$$

where α_w is a weighting constant set to 0.89 empirically. Comparing with the relative threshold offset scheme in [80] and the pure variable relative threshold offset, simulation study shows that the modified relative threshold offset leads to a slight improvement in our speech enhancement approach in terms of both segmental SNR and Itakura Saito measures.

With the above mentioned modified relative threshold offset, we compute the masking threshold. Fig. 2.8 gives the flow chart of the computation of masking threshold, $\Psi(C_{bark})$.

Figure 2.8: Masking threshold computation.

2.4 Methodology of Performance Evaluation

Performance evaluation is usually used to shed light on effectiveness, strengths and weaknesses of speech enhancement systems. The performance of an enhancement system may vary depending on the particular condition in which it is operating. The quality of a speech signal is subjective and reflects the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required by the listeners in order to understand the message. Intelligibility, on the other hand, is an objective measure of the amount of information which can be extracted by listeners from the given signal, whether the signal is clean

or noisy. A given signal may be of high quality and low intelligibility, and vice versa. Hence, these two measures may be independent of each other. Both the quality and the intelligibility of a set of given signals are evaluated based on tests performed on human listeners. Since no mathematical quantification of these measures, in terms of closed-form perceptually meaningful distortion measures, is known, algorithms for the purpose of measuring perceptual improvement and coding robustness are difficult to design and evaluate. Some enhancement systems which improve speech quality may degrade speech intelligibility by using a certain parameter setting. Some systems which improve speech quality may not be applicable if the degradation is due to a competing speaker. As a result, we can only evaluate a speech enhancement method for certain condition and application.

2.4.1 Subjective Measure

It is well known that subjective listening tests are regarded as very important and indispensable for the evaluation of the performance of speech enhancement systems. Speech signals can be compared and evaluated in terms of intelligibility, naturalness, and subjective attributes. However the subjective nature of human listening must be appropriately managed, and very systematic and proven procedures such as Diagnostic Rhyme Test (DRT), Diagnostic Acceptability Measure (DAM) and Mean Opinion Score (MOS) are usually quite expensive to administer. Amongst them, the Mean Opinion Score (MOS) is one of the most popular and widely accepted subjective measures of “acceptability” that is influenced by the intelligibility and quality of the utterance being evaluated. In this thesis, we only carry out limited subjective tests for certain methods for performance comparison. For most cases, we only use informal listening tests to ascertain the strengths of some of the enhancement methods.

Mean Opinion Score

The Mean Opinion Score (MOS) is used widely in speech coding performance evaluations. It involves an Absolute Category Rating (ACR) procedure. Listeners are required to score the overall quality of a speech utterance being tested without the reference of the original speech

Table 2.1: MOS and corresponding speech quality

<i>SCORE</i>	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

utterance. The MOS score of a speech sample is simply the average of the scores from different listeners. There are five categories as indicated in Table 2.1. As the listeners' subjective preference may vary widely, there is usually a bias associated with a listener. The bias could be neutralized by using a large number of listeners. Usually, the listeners are chosen to be "speech technology-naive". At least 40 listeners are recommended to give reliable MOS scores [ITU-T Recommendation P.800, 1996] in a formal setting.

2.4.2 Objective Measure

Objective measures of speech quality are intended to quantify the underlying distortion in an enhanced speech signal. Finding a good objective speech quality measure that correlates well with the subjective judgement of speech quality is important in speech enhancement. Compared to the subjective quality measures, objective measures are more reproducible and less expensive to administer. They are commonly based on the average segmental Signal-to-Noise Ratio (SNR) or some metric distance between the clean speech and the enhanced speech calculated by means of spectral coefficients, linear predictive coefficients (LPC), cepstral coefficients, etc.

Segmental SNR measure

It is well known that the traditional global signal-to-noise ratio (SNR) measure has little perceptual meaning and is thus not a good measure of a speech enhancement performance. Instead, we use the average segmental SNR measure to provide an objective indication of the performance of a speech enhancement algorithm. The segmental SNR is obtained by computing the SNR measure for each speech frame and averaging these measures over the entire signal. It is more

Figure 2.9: An example of segmental SNR variation with time; for noisy speech with input average segmental SNR = -6.22 dB

reliable as an objective distortion measure than the global SNR computed for an entire speech signal. It is defined as

$$\Xi_{seg} = \frac{1}{N_F} \sum_{l=0}^{N_F-1} \Xi_{frame}(l), \quad (2.27)$$

where N_F is the number of frames and $\Xi_{frame}(l)$ is the frame SNR which is defined as

$$\Xi_{frame}(l) = 10 \log_{10} \left(\frac{\sum_{j=0}^{M-1} s^2(Ml - j)}{\sum_{j=0}^{M-1} [\hat{s}(Ml - j) - s(Ml - j)]^2} \right), \quad (2.28)$$

where s denotes the clean speech signal, $\hat{s}$ the noisy speech or enhanced speech, and M the number of samples per frame, which is typically set to 128 for 8 kHz sampling rate. Fig. 2.9 shows an example of the frame SNR curves and the average segmental SNR for a noisy speech.

Itakura-Saito Distortion Measure

In speech quality assessment, several approaches to develop an objective measure based on a speech model have been reported. The Itakura-Saito (IS) distortion measure is one such measure. The IS measure uses LPC coefficients to measure the distortion of an enhanced speech record. It indicates the ratio of the residual energies produced by the original speech when inverse-filtered using the LPC coefficients derived from the original and estimated (or noisy) speech, respectively. It is very sensitive to spectral mismatch in the formant locations and is less affected by the mismatch in the spectral valleys. This is desirable as the human auditory system is more sensitive to errors in formant frequency and bandwidth. The IS measure is given as follows:

$$\text{IS}(l) = \frac{\sigma_s^2}{\sigma_{\hat{s}}^2} \frac{\hat{\mathbf{a}}(l)^T R_s(l) \hat{\mathbf{a}}(l)}{\mathbf{a}(l)^T R_s(l) \mathbf{a}(l)} - \ln \frac{\sigma_s^2}{\sigma_{\hat{s}}^2} - 1 \quad (2.29)$$

Figure 2.10: An example of IS computation for a noisy speech with average IS = 3.13.

where $\mathbf{a}(l)$ and $\hat{\mathbf{a}}(l)$ represent the vectors of LPC coefficients of clean speech and distorted speech respectively of frame l , and $R_s(l)$ is the autocorrelation matrix of the clean speech of frame l . σ_s^2 and $\sigma_{\hat{s}}^2$ are the all-pole gains squared or energies of one step prediction errors of the clean speech frame and the distorted speech frame respectively. Therefore the IS distortion for a speech utterance, IS_{aver} , is given by the average of all IS values over all the frames of the utterance,

$$IS_{aver} = \frac{1}{N_F} \sum_{l=0}^{N_F-1} IS(l), \quad (2.30)$$

Fig. 2.10 shows an example of the IS curves and the average IS value for a noisy speech signal. The frame length was set to 240 samples with an overlap of 180 samples and a sampling rate of 8kHz. In practice, the IS distortion is obtained by averaging the smallest 85% ~ 95% of IS values over all the frames of the speech signal under examination.

Modified Bark Spectral Distortion Measure

The Bark spectral distortion (BSD) measure considers the Bark frequency feature of the human auditory system to measure the underlying distortion [104]. It is based on the assumption that speech quality is directly related to speech loudness, which is defined as the magnitude of auditory sensation. However the original BSD measure does not consider the masking effect but calculates the distortion level by simply considering the loudness difference between the original and the distorted speech, even when the noise spectrum of the corrupted speech is below the masking threshold. The modified Bark spectral distortion (MBSD) measure [111] incorporates a psychoacoustic model into the BSD. It estimates speech distortion in the loudness domain by

including only audible distortions above the masking threshold in the calculation of the distortion measure. The loudness of the masking threshold is compared to the loudness difference of original and the distorted speech to determine if the distortion is perceptible. If the difference is below the loudness of the masking threshold, the distortion is not perceptible, and the indicator of perceptible distortion $M(i)$ is set to 0. Otherwise, it is set to 1. The calculation of the MBSD measure is given by the following equation:

$$\text{MBSD} = \frac{1}{N_F} \sum_{l=0}^{N_F-1} \left[\sum_{C_{bark}=1}^{B_{bark}} M(C_{bark}) |L_s^{(l)}(C_{bark}) - L_{\hat{s}}^{(l)}(C_{bark})| \right] \quad (2.31)$$

where B_{bark} is the number of critical bands, $M(C_{bark})$ is the indicator of perceptible distortion at the C_{bark} -th critical band, and $L_s^{(l)}$ and $L_{\hat{s}}^{(l)}$ are the loudness vectors of the l -th frame of the original and distorted speech respectively.

Perceptual Evaluation of Speech Quality

Determining the subjective speech quality of a transmission system has always been an expensive and laborious process. The ITU-T Rec. P.862 Perceptual Evaluation of Speech Quality (PESQ) [83] is an objective measurement that predicts the results of subjective listening tests on telephony systems. The resulting quality score is analogous to the subjective “Mean Opinion Score” (MOS) measured using panel tests according to ITU-T P.800 [82]. Its scores are calibrated using a large database of subjective tests. PESQ algorithm is designed to analyze specific parameters of audio, including time warping, variable delays and noise. It is primarily intended for applications in codec evaluation and network testing. The algorithm returns a score from -0.5 to 4.5, with higher score indicating better quality. Recently, PESQ is also used for performance evaluation of speech enhancement systems, as it seems to be applicable to any type of audio system and be able to give a useful indication of the “quality” of the processed speech. However, PESQ does not take into account loudness that affect the perceived quality of a speech enhancement system.

2.5 Database for Performance Assessment

In order to evaluate the performance of speech enhancement methods, speech and noise databases are required. In this connection, the publicly available TIMIT speech database [45] is used as the

speech resource for speech enhancement performance evaluation. Also, a realistically recorded speech database which consists of 3695 name-speech waveforms spoken by 5 speakers with 739 names is used to evaluate speech enhancement performance for speech recognition. NOISEX-92 noise database [2] is used as the noise resource in our simulation study. In addition, real life recorded car noise signals captured under a variety of conditions such as different car speeds, window close/open, different types of road surface (highway/roadway), and passing/overtaking vehicle, are also used to test the performance of speech enhancement methods applied to mobile communication in a moving vehicle.

Chapter 3

Spectral Suppression Method

3.1 Introduction

Spectral suppression is one of the simplest and most popular approaches for single channel speech enhancement in the past three decades. This is because the underlying principle to obtain the estimate of the short-time spectral amplitude (STSA) of a speech signal from a noisy speech observation is straightforward to implement. Most of the STSA estimators ignore the estimation of the phase of a speech signal as it has been well demonstrated that the human ear is not very sensitive to the phase of the speech signal [85]. The major contributions in this area include Wiener filtering [48], two-dimension Wiener filtering [36], spectral subtraction [89, 98], Ephraim and Malah (E-M) minimum mean-square error (MMSE) estimation [115] and log spectral amplitude (LSA) estimation [118]. In connection with spectral suppression, masking properties [15, 80] and speech-presence uncertainty [13, 37] have been studied. Though research on spectral suppression has been going on for a long time, issues such as distortion to the original speech signal and residual noise in the form of musical tones created by the enhancement algorithms remain unsolved.

J. S. Lim investigated the estimation of the short-time spectral magnitudes, $|S_k|$, of a speech signal by using a generalized spectral subtraction expression $[|X_k|^\beta - \mathbf{E}\{|N_k|^\beta\}]^{1/\beta}$ for some constant β [47], where k is the frequency bin index, S_k , X_k and N_k are the discrete Fourier transforms of a windowed segment of speech, noisy speech and noise, respectively, and $\mathbf{E}\{|N_k|^\beta\}$ denotes the expectation of $|N_k|^\beta$. The results of his study indicate that the generalized spectral subtraction

algorithm does not lead to improvements in speech intelligibility at the SNR values considered in his study when the degrading source is additive white noise. However, subjectively, the speech processed by Lim's algorithm with some constant β values is "less noisy". Therefore, the generalized spectral subtraction approach may be useful for those applications where perception of noise reduction, without any significant drop in speech intelligibility, is desired [47]. Furthermore, from careful observation of Lim's simulation results, we notice that the intelligibility of the processed speech improves with the increase in the value of β . In essence, the generalized spectral subtraction method hinges on the estimation of the value of a spectral amplitude/power term based on its expected value. For example, when $\beta=1$, the method defaults to amplitude spectral subtraction which directly uses $\mathbf{E}\{|N_k|\}$ to replace $|N_k|$; when $\beta=2$, it is a power spectral subtraction which not only directly uses $\mathbf{E}\{|N_k|^2\}$ to replace $|N_k|^2$, but also uses $\mathbf{E}\{S_k N_k^*\}$ and $\mathbf{E}\{N_k S_k^*\}$ to replace $S_k N_k^*$ and $N_k S_k^*$. The expectations of $S_k N_k^*$ and $N_k S_k^*$ are equal to zero due to the statistical independence assumption [48], where N_k^* and S_k^* represent complex conjugates of N_k and S_k . For $\beta=2$, the case is equivalent to estimating the square root of the maximum likelihood estimator of each signal spectral component variance based on a complex Gaussian model [89]. However, generalized spectral subtraction is not optimal for spectral amplitude estimation under the assumed statistical model and criterion [115]. Consequently, optimal estimation of a speech signal cannot be obtained based on the mathematical model of generalized spectral subtraction, which generally leads to no improvement in the intelligibility of the processed speech. Another study on the effectiveness of the β -order method based on the spectral noise subtraction method can be found in [68]. It also shows a great reduction in the background noise with very little improvement in the intelligibility of speech. The presence of obvious and annoying musical tones in the processed speech caused by the imperfect model of generalized spectral subtraction is yet another one of its drawbacks.

In contrast, the E-M probability-based MMSE noise suppression method [115, 118] is an optimal estimation method as its solution is derived mathematically by minimizing the mean-square error cost function based on the Gaussian model and the statistical independence assumption.

In [81], the elimination of musical noise phenomenon of the E-M suppression method is analyzed; it proves that the E-M noise suppressor is effective if a nonlinear smoothing procedure is used to obtain more consistent estimates of the *a priori* and *a posteriori* SNRs which are used to control the gain function. The advantage of the E-M noise suppression method arises from the non-linearity of the averaging procedure, which leads to a very significant reduction of the musical tone phenomenon [81]. Therefore, the total effect of the E-M noise suppression method is improved as compared to other conventional spectral suppression methods.

In order to investigate the characteristics and performance of the β -order MMSE method based on the assumption that each individual spectral component of speech and noise is a statistically independent complex Gaussian random variable, the use of the cost function $\mathbb{J} = E\{(A_k^\beta - \hat{A}_k^\beta)^2\}$ as an estimation criterion is considered in this chapter. $\hat{A}_k$ is the estimate of spectral amplitude of the speech signal whose spectral component is $S_k = A_k e^{j\alpha_k}$.

In Section 3.2, we describe the β -order STSA MMSE estimation. In order to obtain a more accurate estimate of the speech signal and achieve sufficient suppression of noise and minimal musical tones in the residual signal, a series of β -order MMSE estimators are investigated and their performances are examined. Section 3.3 discusses how the β value could be adapted using the frame signal-to-noise ratio (SNR). To achieve more preservation of weak speech spectral components and more noise reduction, a time-frequency adaptive β -order MMSE method is proposed in Section 3.4. In addition, speech-presence uncertainty is also exploited to achieve additional suppression of noise. In Section 3.5, perceptual auditory masking effect is incorporated into an adaptive β -order MMSE speech enhancement algorithm. The relationship between the value of β and the noise masking threshold is introduced and analyzed. The algorithm is based on a criterion by which the inaudible noise may be masked rather than suppressed. It thereby reduces the chance of introducing distortion to the speech signal by the enhancement algorithm. Moreover, we analyze the musical noise elimination and outline an approach to reduce the computational complexity of the enhancement algorithm in Sections 3.6 and 3.7, respectively. In Section 3.8, the performance of the proposed speech enhancement scheme is evaluated through spectrogram

inspection, objective speech distortion measures, and subjective listening tests using several types of noise sources from the NOISEX-92 database. Section 3.9 gives the summary of this chapter.

3.2 Proposed β -Order MMSE Short-Time Spectral Suppression

3.2.1 Gain Function of β -Order MMSE Estimator

An observed noisy speech signal $x(t)$ is assumed to be a clean speech signal $s(t)$ degraded by uncorrelated additive noise $n(t)$, i.e.,

$$x(t) = s(t) + n(t), \quad 0 \leq t \leq \Delta_T. \quad (3.1)$$

Let $S_k = A_k e^{j\alpha_k}$, N_k , and $X_k = Z_k e^{j\phi_k}$ denote the k th spectral component of the clean speech signal $s(t)$, noise $n(t)$, and the observed noisy speech $x(t)$, respectively, in the analysis interval $[0, \Delta_T]$. We are looking for the estimate of A_k , $\hat{A}_k$, which minimizes the mean-square error between the β -order speech spectral amplitude and the estimated β -order speech spectral amplitude

$$\mathbb{J} = \mathbf{E}\{(A_k^\beta - \hat{A}_k^\beta)^2\} \quad (3.2)$$

given the observed signal $\{x(t), 0 \leq t \leq \Delta_T\}$, where $\mathbf{E}$ indicates the expectation operator. Obviously, the estimator is given by

$$\hat{A}_k = \sqrt[\beta]{\mathbf{E}\{A_k^\beta | x(t), 0 \leq t \leq \Delta_T\}}. \quad (3.3)$$

Assuming the individual spectral components are statistically independent of one another, the expected value of A_k given $\{x(t), 0 \leq t \leq \Delta_T\}$ is equal to the expected value of A_k given X_k only. We therefore have

$$\hat{A}_k = \sqrt[\beta]{\mathbf{E}\{A_k^\beta | X_k\}}. \quad (3.4)$$

By minimizing the cost function $\mathbb{J}$ in Eq. (3.2) with respect to $\hat{A}_k$, the expectation of the β -order A_k given X_k can be derived as

$$\mathbf{E}\{A_k^\beta | X_k\} = \frac{\int_0^\infty \int_0^{2\pi} a_k^\beta p(X_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(X_k | a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k} \quad (3.5)$$

where the symbol a_k denotes the sample value of A_k . With the complex Gaussian assumption of each individual spectral component of speech and noise, the conditional probability density function (PDF) of the observed spectral component given a_k and α_k , $p(X_k|a_k, \alpha_k)$, can be expressed as

$$p(X_k|a_k, \alpha_k) = \frac{1}{\pi\eta_n(k)} \exp\left\{-\frac{|X_k - a_k e^{j\alpha_k}|^2}{\eta_n(k)}\right\} \quad (3.6)$$

and the joint PDF of the speech spectral amplitude and phase, $p(a_k, \alpha_k)$, is given by

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\eta_s(k)} \exp\left\{-\frac{a_k^2}{\eta_s(k)}\right\} \quad (3.7)$$

where $\eta_n(k) = \mathbf{E}\{|N_k|^2\}$, and $\eta_s(k) = \mathbf{E}\{|S_k|^2\}$ are the variances of the k th spectral components of noise and speech signals, respectively. Substituting Eqs. (3.6) and (3.7) into Eq. (3.5), and using the integral representation of the modified Bessel function of zero order $I_0(\cdot)$ [40], we have

$$\begin{aligned} \mathbf{E}\{A_k^\beta|X_k\} &= \frac{\int_0^\infty \int_0^{2\pi} a_k^\beta \frac{1}{\pi\eta_n(k)} \exp\left\{-\frac{|X_k - a_k e^{j\alpha_k}|^2}{\eta_n(k)}\right\} \frac{a_k}{\pi\eta_s(k)} \exp\left\{-\frac{a_k^2}{\eta_s(k)}\right\} d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} \frac{1}{\pi\eta_n(k)} \exp\left\{-\frac{|X_k - a_k e^{j\alpha_k}|^2}{\eta_n(k)}\right\} \frac{a_k}{\pi\eta_s(k)} \exp\left\{-\frac{a_k^2}{\eta_s(k)}\right\} d\alpha_k da_k} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_k^{\beta+1} \exp\left\{-\frac{a_k^2}{\eta(k)}\right\} \exp\left\{2a_k \sqrt{\frac{v_k}{\eta(k)}} \cos(\alpha_k - \vartheta_k)\right\} d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} a_k \exp\left\{-\frac{a_k^2}{\eta(k)}\right\} \exp\left\{2a_k \sqrt{\frac{v_k}{\eta(k)}} \cos(\alpha_k - \vartheta_k)\right\} d\alpha_k da_k} \\ &= \frac{\int_0^\infty a_k^{\beta+1} \exp\left\{-\frac{a_k^2}{\eta(k)}\right\} I_0\left(2a_k \sqrt{\frac{v_k}{\eta(k)}}\right) da_k}{\int_0^\infty a_k \exp\left\{-\frac{a_k^2}{\eta(k)}\right\} I_0\left(2a_k \sqrt{\frac{v_k}{\eta(k)}}\right) da_k} \quad (3.8) \\ &= \frac{\eta(k)^{-1} \Gamma\left(\frac{\beta}{2} + 1\right) \mathbb{M}\left(\frac{\beta}{2} + 1; 1; v_k\right)}{\eta(k)^{-\beta/2-1} \Gamma(1) \mathbb{M}(1; 1; v_k)} \\ &= \eta(k)^{\beta/2} \Gamma\left(\frac{\beta}{2} + 1\right) \mathbb{M}\left(-\frac{\beta}{2}; 1; -v_k\right). \end{aligned}$$

$\Gamma(h)$ is the gamma function and $\mathbb{M}(\alpha; \gamma; z)$ is the confluent hypergeometric function [40], i.e.,

$$\Gamma(h) = \int_0^\infty t^{h-1} e^{-t} dt \quad (3.9)$$

$$\mathbb{M}(\alpha; \gamma; z) = 1 + \frac{\alpha}{\gamma} \frac{z}{1!} + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)} \frac{z^2}{2!} + \frac{\alpha(\alpha+1)(\alpha+2)}{\gamma(\gamma+1)(\gamma+2)} \frac{z^3}{3!} + \dots \quad (3.10)$$

where $\eta(k)$ and v_k are defined as follows:

$$\eta(k) = \left[\frac{1}{\eta_s(k)} + \frac{1}{\eta_n(k)} \right]^{-1} \quad (3.11)$$

$$v_k = \frac{\xi_k}{1 + \xi_k} \gamma_k. \quad (3.12)$$

ξ_k and γ_k represent the *a priori* SNR and *a posteriori* SNR respectively [115], i.e.,

$$\xi_k = \frac{\eta_s(k)}{\eta_n(k)}, \quad \gamma_k = \frac{Z_k^2}{\eta_n(k)}. \quad (3.13)$$

The estimate of the spectral amplitude of the speech signal is then obtained by

$$\begin{aligned} \hat{A}_k &= [\mathbf{E}\{A_k^\beta | X_k\}]^{1/\beta} \\ &= \eta(k)^{1/2} \left[\Gamma\left(\frac{\beta}{2} + 1\right) \mathbb{M}\left(-\frac{\beta}{2}; 1; -v_k\right) \right]^{1/\beta} \end{aligned} \quad (3.14)$$

and the estimate of speech spectral component is given as follows:

$$\hat{S}_k = G_\beta(\xi_k, \gamma_k) X_k. \quad (3.15)$$

$G_\beta(\xi_k, \gamma_k)$ is the gain function which is given by

$$G_\beta(\xi_k, \gamma_k) = \frac{\sqrt{v_k}}{\gamma_k} \left[\Gamma\left(\frac{\beta}{2} + 1\right) \mathbb{M}\left(-\frac{\beta}{2}; 1; -v_k\right) \right]^{1/\beta}. \quad (3.16)$$

Fig. 3.1 (a) shows the gain curves of the β -order MMSE estimator in comparison with the Wiener estimator as a function of γ_k and ξ_k . Fig. 3.1 (b) shows the gain curves with different β values when ξ_k equals to -5dB. The *a priori* SNR, ξ_k , can be well estimated by the decision-directed approach [115] as follows:

$$\hat{\xi}_k(l) = (1 - \alpha) \max[\gamma_k(l) - 1, 0] + \alpha \frac{|G_\beta(\hat{\xi}_k(l-1), \gamma_k(l-1)) X_k(l-1)|^2}{\eta_n(k)} \quad (3.17)$$

where l denotes the frame index. Normally, the parameter α is set to 0.98 [115, 81].

Figure 3.1: (a) Gain versus instantaneous SNR $(\gamma_k - 1)$ in comparison with the Wiener gain (bold-dashed) for $\beta = 0.001$ (dotted), $\beta = 1.0$ (dotdash), $\beta = 2.0$ (dashed) and $\beta = 4.0$ (solid), and a priori SNR $\xi_k = -15, -5, 15$ dB. (b) Gain versus instantaneous SNR $(\gamma_k - 1)$, for different β values and a priori SNR $\xi_k = -5$ dB.

3.2.2 Characteristic of β -Order STSA MMSE Estimator

From Fig. 3.1 (a), it is noted that the gain value, G_β , always converges to the Wiener gain value if the instantaneous SNR $(\gamma_k - 1)$ is big enough for a certain *a priori* SNR (ξ_k) value regardless of the β value. In Fig. 3.1 (b), we can see that when β approaches 0 (e.g., $\beta = 0.001$), the gain curve is very close to the E-M LSA [118] gain curve. When $\beta = 1$, it is exactly the same as the E-M STSA-MMSE [115] gain curve.

Fig. 3.2 (a) shows the gain curves as a function of β for different ξ_k values when $(\gamma_k - 1) = 0$ dB, and Fig. 3.2 (b) shows the gain curves as a function of β value, for different $(\gamma_k - 1)$ for the case of $\xi_k = 0$ dB. Both figures indicate that gain increases as the value of β increases. Fig. 3.2 (b) also illustrates that when the value of $(\gamma_k - 1)$ decreases, the increment of gain (in dB) increases as β becomes higher, for the case of $\xi_k = 0$ dB. For high enough instantaneous SNR value (Fig. 3.2 (b)), strong speech spectral components will be attenuated by almost the same amount for both high and low β values because the β -order based gain value always converges

to the Wiener gain. However, the weak spectral components (i.e. low (γ_k-1) value), which exist in the same frame, will be appropriately preserved because the gain value is big for the low instantaneous SNR spectral components with high β value. This particular characteristic of gain is very important for the preservation of weak spectral components of a speech signal.

For most conventional speech enhancement methods, when a speech spectral component with high SNR value is suitably enhanced by using the appropriate gain value, the weak speech spectral components will be overly attenuated if the same attenuation is applied to all the spectral components in a frame. In contrast, β -order MMSE can exploit the relation between β and gain by adjusting β to a proper value in order not to over attenuate the weak spectral components in a frame. For a low value of β used on a frame of noisy speech samples, the strong speech spectral components can be appropriately enhanced but the weak speech spectral components will be lost. When the value of β is high, the strong speech spectral components remain at almost the same enhancement level as in the case of low β value because the gain always converges to the Wiener gain value for high enough (γ_k-1) value. However, the weak spectral components (i.e., components with low (γ_k-1) values), which exist in the same frame as the strong speech spectral components, may be appropriately preserved because the gain has a big value for the low instantaneous SNR spectral components when high β value is used.

If a high value of β is used for a speech absence frame, the musical tones will increase. If too low a β value is used in a speech presence frame, the weak speech spectral components will be lost although the musical tones will also be attenuated. If we increase the β value in a speech presence frame, the spectral components with weak SNR will be raised and therefore they could be better estimated, though the noise (musical tones) will be enlarged correspondingly. Fortunately, according to the masking principle [80], the weak musical tones might be masked by the strong speech signal depending on the relative frequency separation between the maskees and the maskers. The effects of a high β value acting on a speech-presence frame and a low β value acting on a speech-absence frame are the salient attributes of the proposed β -order MMSE method.

Figure 3.2: (a) Gain versus β value for *a priori* SNR $\xi_k = -20, -10, 0, 10, 20$ dB and instantaneous SNR $(\gamma_k - 1) = 0$ dB. (b) Gain versus β value for instantaneous SNR $(\gamma_k - 1) = -20, -10, 0, 10, 20$ dB and *a priori* SNR $\xi_k = 0$ dB.

According to O. Cappé's work [81] on E-M MMSE, when the speech level is well above the noise level, the *a priori* SNR estimation equation involves a mere one-frame delay of instantaneous SNR; when the speech signal level is close to or below the noise level, the *a priori* SNR estimation equation has a smoothing effect. Consequently, we can reasonably assume $\xi_k = \gamma_k - 1$ as a meaningful possible path for tracking the gain changes when we discuss different related gain estimators. As shown in Figure 3.3, we can see that Wiener filtering may over-attenuate the weak signals (low SNR), E-M LSA (similar to $\beta = 0.001$ case) performs better in preserving the weak signals, and β -order MMSE has obviously a much wider range of gain values and thus more flexible and effective for estimating weak spectral components provided that the β value can be appropriately adapted for different signal component strengths. The over-attenuation of weak spectral components (i.e., low SNR) in the case of Wiener filtering can be observed clearly through inspection of spectrograms obtained from a large number of simulations using many different speech utterances corrupted by different types of noise.

Figure 3.3: Gain curves for the cases of instantaneous SNR $(\gamma_k - 1)$ equals to ξ_k for the Wiener gain, MMSE with $\beta = 0.001$ and $\beta = 4.0$.

3.2.3 Investigation of MMSE Estimator for Constant β Value

To understand the effect of a constant β value, used in the β -order MMSE estimation of speech spectral amplitudes, on the trade-off between speech distortion and noise reduction, a large number of simulation experiments are carried out by using 30 phonetically-balanced sentences from the TIMIT corpus [45], and 4 types of noise from NOISEX-92 [2]. It is well known that the traditional signal-to-noise ratio (SNR) has little perceptual meaning and is thus not a good measure of speech quality. Instead, the average segmental SNR (seg.SNR) is a relatively better measure, which is obtained by computing the SNR value of each short time frame and averaging the values of all the frames of an entire speech utterance. In our experiments, five different input seg.SNRs [95] (-10, -5, 0, 5, 10 [dB]) and 9 β -order MMSE estimators ($\beta = 0.001, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0$) are used in the investigation. In addition to seg.SNR, we also use PESQ to assess the enhancement performance.

It is well known that the performance of a single channel speech enhancement algorithm is significantly affected by the estimation of noise. Effective and commonly used noise estimation methods are described in [27, 91]. In [68], an over-estimate of noise power spectrum is found to

result in a great reduction in the background noise while maintaining almost the same intelligibility of speech when a spectral subtraction method is used. In our β -order MMSE study, we use the conventional minimum statistics noise tracking method [91] which searches the minimum periodogram of a noisy signal to obtain the shape of noise power spectrum. In general, the power spectral values of the minimum periodogram need a compensation multiplicative factor, which is denoted as α_e , to approach their expected values [91, 92]. We select the factor α_e by using the following method: We compare the minimum power spectral values of the noisy speech with the average power spectral values of the original noise signal to obtain the appropriate α_e for each spectral bin. The different α_e s of all the spectral bins are then averaged to arrive at the final overall α_e which is then fixed for all frames and frequency bins in our simulation for all the different utterances. We refer to the obtained α_e value as the normal compensation factor and the corresponding noise estimation as normal-estimation. We also consider another noise compensation factor which is greater than α_e , and denote it as α_{over} . We call α_{over} the over-compensation factor and noise estimation based on it as noise over-estimation. We give two sets of experimental results which are under the same conditions except that the noise compensation factors are α_e and α_{over} respectively. In our simulation, we set $\alpha_{over} = 1.6\alpha_e$.

Fig. 3.4 shows the performance of the speech estimators based on Eq. (3.16) in terms of average seg.SNR improvement for different β values at 8 kHz sampling rate. Sub-figures (a1) and (b1) show the improvement in average seg.SNR for the cases of white Gaussian noise and F16 cockpit noise respectively, for different input seg.SNRs and using normal noise estimation based on α_e . Similarly, sub-figures (a2) and (b2) are for the same simulation experiments except with the use of the over-compensation factor α_{over} instead of α_e . Fig. 3.5 shows the PESQ performance, instead of improvement in average seg.SNR, of the estimators with the same conditions used in Fig. 3.4.

From these figures, it is found that when the noise power spectrum is obtained by using normal-estimation, the MMSE estimator with low β values always outperforms the estimator with high β values in terms of seg.SNR improvement and PESQ value. However when we overestimate

the noise power spectrum, the use of low β values is not always better than high β values. In addition, it is clear that most of the seg.SNR performance of β -order MMSE with the over-estimation of noise power spectrum is better than the case of normal-estimation. However, the PESQ performances for the two cases seem to be very close. It can also be seen that the PESQ measure of the β -order MMSE processed speech for the two cases of noise estimation is greatly improved as compared to the original noisy speech. Furthermore, the PESQ value becomes higher for higher β values with the change of noise estimation from normal to over-estimation.

3.3 Adaptive β -Order MMSE

As discussed in Section 3.2, the estimate of speech spectral amplitude A_k given $x(t)$ ($0 \leq t \leq T$) is based on modelling speech and noise spectral components as statistically independent Gaussian random variables, which is well elaborated in [115]. The statistical independence assumption in the Gaussian model is equivalent to the assumption that the Fourier expansion coefficients are uncorrelated. As a consequence, we can have $\hat{A}_k = \sqrt[\beta]{\mathbf{E}\{A_k^\beta | x(t), 0 \leq t \leq T\}} = \sqrt[\beta]{\mathbf{E}\{A_k^\beta | X_1, X_2, \dots, X_K\}}$ be equivalent to $\sqrt[\beta]{\mathbf{E}\{A_k^\beta | X_k\}}$, where K denotes the number of frequency bins. Therefore the k -th spectral gain of Eq. (3.16) is only a function of *a priori* and *a posteriori* SNRs of the k -th frequency bin rather than the SNRs of other frequency bins.

The statistical independence assumption is reasonably valid and significant for deriving the expression for $\hat{A}_k$. However, it ignores the statistical information from neighbor frequencies. It is believed that for estimation of a spectral amplitude at certain frequency, besides the statistical information at its own frequency, the statistical information at other frequencies is also useful. Consequently, the value of β used in our estimators is designed to adapt to the frame SNR which takes into account all the spectral components in a frame. This approach has been found to be experimentally effective and sufficient in achieving good simulation results for our estimator. We may arrive at an appropriate β value for a particular frame l by adapting it with respect to frame SNR which is defined as

$$\Xi(l) = 10 \log_{10} \frac{\sum_{k=0}^{N/2} |S_k(l)|^2}{\sum_{k=0}^{N/2} |N_k(l)|^2}. \quad (3.18)$$

Figure 3.4: Segmental SNR comparison for the estimated speech signals obtained by the MMSE estimators with different β values and input noisy speech with seg.SNR (*) -10 dB, ($\diamond$) -5 dB, (x) 0 dB, (o) 5 dB, and (+) 10 dB.

Figure 3.5: PESQ comparison. Solid line for the estimated speech obtained by the MMSE estimators with different β values, and dashed line for input noisy speech with seg.SNR (*) -10 dB, ($\diamond$) -5 dB, (x) 0 dB, (o) 5 dB, and (+) 10 dB.

As we only have the observed noisy speech signal, but do not have the clean speech signal and noise signal, the frame SNR can be approximated by using the following equation

$$\hat{\Xi}(l) = 10 \log_{10} \frac{\sum_{k=0}^{N/2} |Z_k(l) - \sqrt{\eta_n(l, k)}|^2}{\sum_{k=0}^{N/2} \eta_n(l, k)}. \quad (3.19)$$

In other words, we can make β a function of frame SNR $\Xi(l)$, i.e., $\beta(l) = \mathbb{F}(\Xi(l))$, where $\mathbb{F}$ denotes function operator.

Based on the preceding analysis of the characteristics of β -order MMSE, it is desirable for β to increase in value as $\Xi(l)$ increases, and to decrease in value when $\Xi(l)$ decreases. Here a linear relationship between β and $\Xi(l)$ is applied as follows:

$$\beta(l) = \tau_1 \Xi(l) + \tau_2 \quad (3.20)$$

where τ_1 and τ_2 denote linear coefficients.

In contrast to the conventional optimal estimation approach that is based on a cost function and a mathematical derivation, Eq. (3.20) is an empirical formula with parameters (τ_i) obtained from simulation experiments. Nevertheless, as will be shown in Section 3.8, good simulation results are obtained by setting proper values of these parameters.

In addition to adapting β based on frame SNR, we also limit β to a maximum value of 4 as too high of a β value will leave most of the noise components intact. The final β value is therefore given by

$$\beta(l) = \max\{\min[\tau_1 \Xi(l) + \tau_2, \tau_3], \tau_4\}. \quad (3.21)$$

In our simulation, we empirically set $\tau_1=0.25$, $\tau_2=1.75$, $\tau_3=4.00$, $\tau_4=0.001$. It is found experimentally that the proposed approach can help to preserve the weak speech spectral components in a strong speech-presence frame. This is mainly due to the use of high β values for weak spectral components in this particular situation. For weak speech-presence frames, especially those with weak unvoiced speech, the underlying speech is attenuated as much as noise. In this case, the proposed frame SNR based method is not superior to the conventional E-M MMSE [115] method in improving the intelligibility of the enhanced speech. When the noise completely covers the very weak unvoiced speech, it is very unlikely for any method to be able to recover the

underlying speech. However, the low value of β used in this case would at least suppress most of the annoying noise signal.

As a heuristic method, the adaptation of β depending on frame SNR has its limitation. However, the superior performance of the adaptive β -order MMSE speech enhancement method shows the significance and validity of β -order MMSE. It is envisaged that better performance could be achieved if a more appropriate parameter is used to adapt the β value. One such parameter is the masking threshold which has been well exploited in generalized spectral subtraction for speech enhancement [80, 15]. Another parameter that could enhance the effectiveness of β adaptation is the local SNR besides frame SNR.

3.4 A Time-Frequency Adaptive β -Order MMSE Algorithm with Speech-Presence Uncertainty

3.4.1 Time-Frequency Adaption of β

From Eq. (3.16), we can see that gain increases as the value of β increases, and the smaller the value of the instantaneous SNR ($\gamma_k - 1$) is, the bigger the increment of gain (in dB) will be as β increases. This particular characteristic of gain is very important to the preservation of the weak speech spectral components. For a low value of β used on a frame of noisy speech samples, the strong speech spectral components can be appropriately enhanced but the weak speech spectral components will be lost. When the value of β is high, the strong speech spectral components remain at almost the same enhancement level as in the case of low β value because the gain always converges to the Wiener gain value when the instantaneous SNR is high enough. However, the weak spectral components (i.e. low instantaneous SNR, $(\gamma_k - 1)$) which exist in the same frame as strong spectral components of the speech signal may be appropriately preserved because the gain has a big value for the low instantaneous SNR spectral components with a high β value.

Since the suppression gain function (Eq. (3.16)) is derived based on the statistical independence assumption of both speech and noise spectral components, the estimate of a speech spectral amplitude of a certain frequency bin is only determined by the *a priori* and *a posteriori* SNRs

of that frequency bin, and is independent of the SNRs of the other frequency bins. However, it is well-known that a speech signal is not a chaotic signal and the statistical independence assumption is not sufficiently satisfied in practice. Based on the above analysis, we can make β a function of frame SNR $\Xi(l)$ and local SNR $\Lambda(l, k)$, i.e., $\beta(l, k) = \mathbb{F}(\Xi(l), \Lambda(l, k))$. The local SNR of the current frame l is defined by

$$\Lambda(l, k) = 10 \log_{10} \frac{\sum_{i=-b}^b \sum_{j=-a}^d w(i, j) \varsigma(l - j, k - i)^2}{\sum_{i=-b}^b \sum_{j=-a}^d w(i, j) \eta_n(l - j, k - i)} \quad (3.22)$$

where w is a time-frequency window with size $(2b + 1) \times (a + d + 1)$, d and a denote the numbers of time-domain past and future samples used for smoothing for frequency bin k at time l , and $\varsigma(l, k)$ is defined as follows

$$\varsigma(l, k) = \max[Z_k(l) - \sqrt{\eta_n(l, k)}, \quad \varepsilon]. \quad (3.23)$$

In practice, $b = 1$, $a = 1$, $d = 2$ and matrix w is given by

$$w = \begin{bmatrix} 0.35 & 0.5 & 0.7 & 0.5 \\ 0.50 & 0.8 & 1.0 & 0.8 \\ 0.35 & 0.5 & 0.7 & 0.5 \end{bmatrix}. \quad (3.24)$$

Since local SNR represents the time-frequency SNR, the β value is regulated by the time-frequency information. The β value may be expressed as a function of the two variables in the polynomial form, i.e.,

$$\begin{aligned} \beta(l, k) &= \mathbb{F}(\Xi(l), \Lambda(l, k)) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{ij} \Xi(l)^i \Lambda(l, k)^j \\ &= \tau_0 + \tau_1 \Xi(l) + \tau_2 \Lambda(l, k) + \tau_3 \Xi(l) \Lambda(l, k) + \Phi(O_h) \end{aligned} \quad (3.25)$$

where c_{ij} and τ_i ($i, j = 0, 1, 2, \dots, \infty$) indicate the polynomial coefficients. By ignoring the high-order polynomial terms $\Phi(O_h)$, and making β a monotonically non-decreasing function of $\Xi(l)$ and $\Lambda(l, k)$, we can express $\beta(l, k)$ approximately in the following form

$$\begin{aligned} \beta(l, k) &\cong \tau_0 + \tau_1 \Xi(l) + \tau_2 \max[\Lambda(l, k) - \tau_3, 0] \\ &\quad + \tau_4 \max[\Xi(l) - \tau_5, 0] \max[\Lambda(l, k) - \tau_6, 0]. \end{aligned} \quad (3.26)$$

Firstly, we consider the preservation of the weak spectral components which are contained within the same frame as the strong spectral components. In this case, frame SNR is high, $\Xi(l)$ determines the value of β resulting in a high β value. The weak spectral components inside the frame are thus not over-attenuated. Secondly, when the frame SNR is low, the β value is consequently also low. In this case, most spectral components are sufficiently attenuated so that the strong noise spectral components are appropriately suppressed. At the same time, we let local SNR have a significant weight so that the perceivable spectral components with sufficiently high local SNR are not lost. Thirdly, in the case of a speech pause with very low frame SNR, we let $\beta \rightarrow 0$ to attenuate almost all the noise energy. From Eq. (3.26), we can see that the parameters τ_i ($i = 0, 1, \dots, 6$) are very important to the speech enhancement system. How to correctly obtain the parameters is a key issue in our adaptive β -order design. Actually, τ_0 is a floor to determine the total level of β value, and τ_1 and τ_2 determine to what extent the frame and local SNRs influence the suppression gain value respectively. τ_4 is used to constrain the overall effect of frame and local SNRs on the β value. τ_3 , τ_5 and τ_6 are to set a lower bound contributions to the gain function from frame and local SNRs. To emphasize the dynamic range of the β value, we express the final form of β as follows

$$\hat{\beta}(l, k) = \min \{ \max\{\beta(l, k), \tau_7\}, \tau_8 \}. \quad (3.27)$$

Through a large number of computer simulation using real speech data, the τ_i ($i = 0, 1, 2, \dots$) parameters are appropriately adjusted as follows: $\tau_0 = 0.95$, $\tau_1 = 0.12$, $\tau_2 = 0.09 \sim 0.1$, $\tau_3 = -5 \sim -1$ (dB), $\tau_4 = 0.015$, $\tau_5 = -6.7$ (dB), $\tau_6 = -1.2$ (dB), $\tau_7 = 0$ and $\tau_8 = 4$.

3.4.2 Suppression Gain Incorporating Speech-Presence Uncertainty Gain

Assuming $\mathbb{H}_0$ and $\mathbb{H}_1$ indicate speech-absence and speech-presence respectively in a binary hypothesis model, the probability density functions (pdfs) of $X_k(l)$ with Gaussian distribution during speech-absence and speech-presence are respectively given by

$$p(X_k(l)|\mathbb{H}_0(k, l)) = \frac{1}{\pi\eta_n(k, l)} \exp\left[-\frac{|X_k(l)|^2}{\eta_n(k, l)}\right] \quad (3.28)$$

$$\begin{aligned}
& p(X_k(l)|\mathbb{H}_1(k, l)) \\
&= \frac{1}{\pi(\eta_n(k, l) + \eta_s(k, l))} \exp\left[-\frac{|X_k(l)|^2}{\eta_n(k, l) + \eta_s(k, l)}\right].
\end{aligned} \tag{3.29}$$

According to the Bayes' theorem, we have

$$\frac{p(\mathbb{H}_1(k, l)|X_k(l))p(X_k(l))}{p(\mathbb{H}_0(k, l)|X_k(l))p(X_k(l))} = \frac{p(X_k(l)|\mathbb{H}_1(k, l))p(\mathbb{H}_1(k, l))}{p(X_k(l)|\mathbb{H}_0(k, l))p(\mathbb{H}_0(k, l))}. \tag{3.30}$$

As $p(\mathbb{H}_1(k, l)|X_k(l)) + p(\mathbb{H}_0(k, l)|X_k(l)) = 1$, we have

$$p(\mathbb{H}_1(k, l)|X_k(l)) = \frac{1 - \varrho_k(l)}{1 - \varrho_k(l) + \varrho_k(l)(1 + \xi_k)\exp(-v_k)} \tag{3.31}$$

where $\varrho_k(l) = p(\mathbb{H}_0(k, l))$, $p(\mathbb{H}_1(k, l)) = 1 - \varrho_k(l)$, and l is the frame index. The estimate of speech amplitude is then obtained by

$$\begin{aligned}
\hat{A}_k^\beta &= \mathbf{E}\{A_k^\beta|X_k(l), \mathbb{H}_1(k, l)\}p(\mathbb{H}_1(k, l)|X_k) \\
&\quad + \mathbf{E}\{A_k^\beta|X_k(l), \mathbb{H}_0(k, l)\}p(\mathbb{H}_0(k, l)|X_k).
\end{aligned} \tag{3.32}$$

As illustrated in [115], the second term is usually equal to 0. Thus the estimate of speech amplitude may be approximated by

$$\hat{A}_k = G_\beta(\xi_k, \gamma_k)p(\mathbb{H}_1(k, l)|X_k)^{1/\beta}Z_k. \tag{3.33}$$

Taking into account the factor of uncertainty of speech-presence, the suppression gain is given by

$$\begin{aligned}
G_M(\xi_k, \gamma_k, \beta) &= \frac{\sqrt{v_k}}{\gamma_k} [\Gamma(\beta/2 + 1)\mathbb{M}(-\beta/2; 1; -v_k) \\
&\quad \times \frac{1 - \varrho_k(l)}{1 - \varrho_k(l) + \varrho_k(l)(1 + \xi_k)\exp(-v_k)}]^{1/\beta}.
\end{aligned} \tag{3.34}$$

Subsequently, the estimate of a speech spectral component is expressed as $\hat{S}_k = G_M(\xi_k, \gamma_k, \beta)X_k$.

Estimation of the Speech-Absence Probability

In [115], $\varrho_k(l)$ is fixed to a value of 0.2. A more accurate estimate of $\varrho_k(l)$ is to make it adapt to the time-varying speech signal. Consequently, the estimate of the speech signal will be closer to its original value. We therefore adopt the following procedure to estimate $\varrho_k(l)$:

$$p_0(k, l) = (1 - \alpha_p)p_0(k, l - 1) + \alpha_p\Omega(k, l) \tag{3.35}$$

$$\Omega(k, l) = \begin{cases} 0, & \text{if } \gamma_k - 1 > \hbar_p \\ 1, & \text{if } \gamma_k - 1 \leq \hbar_p \end{cases} \quad (3.36)$$

$$\varrho_k(l) = 0.15 + 0.15p_0(k, l) \quad (3.37)$$

In our simulation, $\alpha_p = 0.85$ and $\hbar_p = 1.0 \sim 2.0$ are chosen experimentally.

3.5 Proposed β -Order MMSE with Masking Properties

3.5.1 Gain of a Masking-Based β -Order MMSE Method

With masking-based speech enhancement, the amount of noise to be reduced is set at the level of noise that can be masked by the clean speech in order to reduce the distortion introduced to speech due to the enhancement process. Subsequently, the amount of noise reduction obtained by using a masking-based enhancer during speech activity period may be more than that obtained by using the same enhancer without masking consideration. On the other hand, the amount of noise reduction is not affected during speech pause period. In order to avoid a sudden change from speech activity to speech pause in the perceptual sense, spectral flooring is adopted. We therefore modify the gain function as follows:

$$G_M(\xi_k, \gamma_k, \beta) = \begin{cases} \frac{\sqrt{v_k}}{\gamma_k} [\Gamma(\frac{\beta}{2} + 1) M(-\frac{\beta}{2}; 1; -v_k)]^{1/\beta}, & \text{if } \gamma_k > \rho_1(k) + \rho_2(k); \\ \{\rho_2(k) \frac{1}{\gamma_k}\}^{1/2}, & \text{otherwise,} \end{cases} \quad (3.38)$$

where $\rho_1(k)$ is a threshold factor. Together with $\rho_2(k)$, it determines the lowest SNR limitation below which a spectral floor gain function is applied in place of β -order MMSE gain. $\rho_2(k)$ is a coefficient of spectral floor which is affected by the masking threshold and its minimum and maximum settings. In the case of speech pause period, as the masking threshold is equal to the absolute masking threshold value which is too low to be considered, $\rho_2(k)$ is therefore equal to its minimum value. Through a large number of simulations, we propose that the two parameters be made directly proportional to the masking threshold with the set of minimum and maximum

values which vary from frame to frame ¹, i.e.,

$$\rho_1(k) = \frac{T_f(l, k) - T_{min}(l)}{T_{max}(l) - T_{min}(l)}(\rho_{1max} - \rho_{1min}) + \rho_{1min}, \quad (3.39)$$

$$\rho_2(k) = \frac{T_f(l, k) - T_{min}(l)}{T_{max}(l) - T_{min}(l)}(\rho_{2max} - \rho_{2min}) + \rho_{2min}. \quad (3.40)$$

In the above equations, ρ_{1min} , ρ_{2min} and ρ_{1max} , ρ_{2max} denote the minima and maxima of the parameters ρ_1 and ρ_2 respectively; and $T_{max}(l)$ and $T_{min}(l)$ are the maximum and minimum values of noise masking threshold $T_f(l, k)$ at current frame l . In our simulation, we empirically select $\rho_{1min}=1$, $\rho_{1max}=6.28$, $\rho_{2min}=0$ and $\rho_{2max}=0.015$. As mentioned in [80], spectral flooring leads a reduction of residual noise and an increased level of background noise. The estimate of speech signal is therefore given by

$$\hat{S}_k = G_M(\xi_k, \gamma_k, \beta)X_k. \quad (3.41)$$

3.5.2 Adaptation of β Value Based on Masking Properties

It is well known that some audio spectral components are inaudible to human listeners because they are masked by other components [80, 15, 55, 93, 71]. This implies that human listeners cannot notice the difference between the original speech and the speech distorted by a processing step if the distortions in the processed speech are masked by select original speech spectral components. This auditory masking phenomenon is well exploited for discarding the inaudible audio components or masking quantization noise in audio and speech coding. However, speech enhancement is used not only to improve the perceptual quality of speech but also to improve the recognition rate of an ensuing speech recognition system. Though a speech signal with some inaudible components has almost the same perceptual quality as the one without the inaudible components to human listeners, it may have different effects when it is input to a speech recognition system where some weak spectral components, though inaudible, may be critical to the successful recognition of some phonemes. The results reported in [80] substantially confirm this conjecture.

¹In [80], combining with the generalized spectral subtraction algorithm, they are reciprocally proportional to the masking threshold with different set of maximum and minimum values.

The maximum allowable noise spectrum (or distortion spectrum) which is not perceivable by a human listener is called the noise masking threshold. Using noise masking threshold as weighting factor but not as a threshold has been applied to speech enhancement [80]. We can use auditory masking effects to achieve the classic trade-off between speech distortion and noise reduction, wherein the inaudible noise spectral components together with the corresponding speech spectral components are not attenuated, and thereby reducing the chance of further distortion to speech. On the other hand, those noise spectral components that are audible must be removed [15].

We propose to make the β value such that it not only depends on the frame SNR which represents the power of the speech signal in the current frame but also relies on the noise masking threshold in a particular Bark band, in order to achieve an optimal trade-off between speech distortion and noise reduction. According to the Taylor polynomial, we can express β as a function of two variables in the following form

$$\begin{aligned}\beta(l, k) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{ij} \Xi(l)^i \Theta_f(l, k)^j \\ &= \tau_0 + \tau_1 \Xi(l) + \tau_2 \Theta_f(l, k) + \tau_3 \Xi(l) \Theta_f(l, k) + \Phi(O_h),\end{aligned}\tag{3.42}$$

where Θ_f represents the perceptual factor (to be discussed later in the paper) of the human auditory system in the frequency domain, c_{ij} and τ_i ($i, j = 0, 1, 2, \dots, \infty$) denote the polynomial coefficients, and $\Phi(O_h)$ represents the high-order polynomial terms. By ignoring the high-order polynomial terms, and assuming β is a monotonic non-decreasing function of $\Xi(l)$, we can express $\beta(l, k)$ approximately in the following form

$$\hat{\beta}(l, k) \cong \tau_0 + \tau_1 \Xi(l) + \tau_2 \Theta_f(l, k) + \tau_3 \max[\Xi(l) - \tau_4, 0] \Theta_f(l, k).\tag{3.43}$$

From Eq. (3.43), we can see the parameters τ_i ($i = 0, 1, \dots, 4$) determine the effectiveness of the β -order MMSE speech enhancement system. Actually, τ_0 is a floor that determines the general level of β value, and τ_1 and τ_2 decide the extent the frame SNR and the masking threshold influence the suppression gain value. τ_3 adds another contribution that is based on both the frame SNR and noise masking threshold, while τ_4 is to give a lower bound contribution to gain function from frame SNR. It means frame SNRs below τ_4 will not affect the gain value.

Through a large number of computer simulation experiments using real speech and noise data, the τ_i ($i = 0, 1, 2, \dots$) parameters are appropriately adjusted. Here, we give the empirically obtained polynomial coefficients of Eq. (3.43) as follows

$$\hat{\beta}(l, k) = 0.942 + 0.121\Xi(l) + 0.981\Theta_f(l, k) + 0.187 \max[\Xi(l) + 6.7, 0]\Theta_f(l, k). \quad (3.44)$$

In contrast to the conventional optimal estimation approach based on a cost function and using mathematical derivation, Eq. (3.44) is an empirical formula with parameters (τ_i) obtained from simulation experiments. As will be shown in Section 3.8, good simulation results are obtained by using these parameter values. Further justifications for Eq. (3.44) are given in the following paragraph.

In our proposed β -order MMSE based speech enhancement, a higher masking threshold will result in a higher β value and consequently a higher gain in the estimation process. This will ensure that some speech spectral components would not be over-attenuated and thus minimizes distortions to the speech signal. Based on the noise masking threshold, we can determine the power and spectral shape of noise that might be inaudibly inserted into the speech signal. Consequently, the frequency-dependent masking threshold can be regarded as the reference spectrum for re-shaping the noise. If the masking threshold is high, residual noise will mostly be masked and become inaudible. Hence, there is no need to attenuate it in order not to increase the level of distortion to speech. In this case, the β value should be high which results in a relatively high gain value for weak spectral components. However, if the masking threshold is low, residual noise will be annoying to human listeners and it has to be reduced in amplitude. This is done by reducing the β value appropriately. In short, we should use high β value for components with high SNR values, and also when the masking threshold is high.

In addition to adapting β based on frame SNR and masking threshold and using Eq. (3.44), we also limit β to a maximum value of 4 as too high a β value will leave most of the noise components intact. The final β value is therefore given by

$$\tilde{\beta}(l, k) = \min \{ \max\{\hat{\beta}(l, k), 0.001\}, 4 \}. \quad (3.45)$$

The perceptual factor Θ is obtained by the following normalization step

$$\Theta(l, \lambda) = \frac{T(l, \lambda) - T_{\min}(l)}{T_{\max}(l) - T_{\min}(l)}. \quad (3.46)$$

The frequency domain perceptual factor Θ_f has its corresponding expression Θ in the Bark domain, i.e., $\Theta_f(l, k) = \Theta(l, \lambda)$.

Fig. 3.6 gives an example of the value of β adaptation by using Eq. (3.44) for a speech utterance contaminated by F-16 cockpit noise with an input segmental SNR of -10 dB.

Just like most masking-based speech enhancement methods, the noise masking threshold $T(l, \lambda)$ has to be obtained from a speech signal. We therefore determine $T(l, \lambda)$ from an estimated speech signal which is obtained using β -order MMSE method with β being a certain fixed value. In our simulation, we set β to 0.5.

3.5.3 Implementation of Masking-based Speech Enhancement Method

In this section, we describe the implementation of the proposed masking-based speech enhancement system. As shown in Fig. 3.7, a noisy speech signal is Hanning windowed and then transformed into the spectral domain. The estimate of noise spectral amplitude is obtained during speech pause period. Following the pre-estimation of speech spectral amplitude by using a fixed-order MMSE method, the masking threshold is computed. With the noisy speech spectral amplitude and the estimate of noise, the frame SNR is provided. Consequently, the value of β is calculated according to the masking threshold and the frame SNR. Finally, the enhancement processing is done by using the masking-based β -order MMSE gain function with the adapted β value.

3.6 Analysis of Musical Tone Elimination

The elimination of musical tones using the E-M STSA-MMSE [115] estimator is described in [81]. It is mainly due to the effectiveness of the so called ‘decision-directed approach’ for estimating the *a priori* SNR, ξ_k . Obviously, this musical tone elimination feature can also be applied to the β -order MMSE described by Eq. (3.16) for any value of β . Figure 3.8 is obtained by using

Figure 3.6: The adaptation of the values of β by using our proposed masking-based adaptation (a) The values of β adapted by using Eq. (3.44) at frequency 1000 Hz and 2000 Hz, for a speech utterance contaminated by F-16 noise with input segmental SNR of -10 dB; (b) The short-term spectral amplitude of the clean speech utterance at frequency 1000 Hz and 2000 Hz; (c) The theoretic frame SNR curve by using Eq. (3.18) and the estimated frame SNR by using Eq. (3.19) and the segmental SNR of -10 dB for the input noisy speech; (d) The waveform of the clean speech utterance.

Figure 3.7: The structure of the proposed masking-based adaptive β -order MMSE method.

the adaptive β -order MMSE method. It shows an example of the changes in *a priori* SNR and instantaneous SNR versus frame number at the frequency bin of 1 kHz. We can see that when the speech level is well above the noise level, the frame-based *a priori* SNR estimation equation involves a mere one-frame delay relative to the instantaneous SNR, and it is not a smoothed SNR estimate, which is more accurate for a non-stationary signal. When the speech signal level is close to or below the noise level, the *a priori* SNR estimation equation has a smoothing property and the musical tone phenomenon is therefore greatly reduced. This is similar to the case of E-M MMSE which has the effect of musical tone elimination [81]. Thus it verifies the capability of the proposed adaptive β -order MMSE method to eliminate musical tones. On the other hand, as the attenuation of the MMSE gain depends mainly on the value of *a priori* SNR [81], constraining *a priori* SNR, ξ_k , to be larger to a threshold (e.g., -15 dB) will further eliminate the musical noise phenomenon.

Figure 3.8: Variations of *a priori* SNR ξ_k and instantaneous SNR $(\gamma_k - 1)$ versus frame number. (k corresponds to 1 kHz.)

3.7 Reduction of Computational Complexity

To implement the adaptive β -order MMSE algorithm for real-time realization, the computational complexity involved in Eq. (3.16) has to be simplified. The value of β as well as v_k (Eq. 3.12) can be discretized instead of being continuous real values, and the values of the confluent hypergeometric function and the gamma function can be obtained by using lookup tables. This can greatly reduce the implementation complexity of the two functions. In our simulation, we use the following simplifications:

Eq. (3.16) can be re-written as

$$G_\beta(\xi_k, \gamma_k) = \Upsilon(\gamma_k) \Phi(\beta, v_k) \quad (3.47)$$

where $\Phi(\beta, v_k) = \sqrt{v_k} [\Gamma(\frac{\beta}{2} + 1) \mathbb{M}(-\frac{\beta}{2}; 1; -v_k)]^{1/\beta}$ and $\Upsilon(\gamma_k) = \frac{1}{\gamma_k}$. It can be seen that $\Phi(\beta, v_k)$ consists of the confluent hypergeometric function and the gamma function, and it has two parameters, i.e., v_k and β . As a consequence, the computation of the gain function (3.16) related to the confluent hypergeometric function and the gamma function is reduced to compute the product of $\Upsilon(\gamma_k)$ and $\Phi(\beta, v_k)$. The latter can be obtained by searching a lookup table. In the lookup table of $\Phi(\beta, v_k)$, the value of $\Phi(\beta, v_k)$ is discretized to $\tilde{\Phi}(\tilde{\beta}, \tilde{v})$.

We constrain $\tilde{v}$ to the range of $[-17 \text{ dB}, 35 \text{ dB}]$ ² with a step-size of 0.1 dB. A total of 521 values of $\tilde{\Phi}(\tilde{\beta}, \tilde{v})$ for each $\tilde{\beta}$ value are tabulated. For the case of $\beta = \tilde{\beta}_m$, estimate of $\Phi(\tilde{\beta}_m, v_k)$ is calculated as follows. We assume a true v_k value falls in between two adjacent values, $\tilde{v}_n$ and $\tilde{v}_{n+1}$, in the table, i.e, $\tilde{v}_n \leq v_k \leq \tilde{v}_{n+1}$. $\tilde{\Phi}(\tilde{\beta}_m, \tilde{v}_n)$ and $\tilde{\Phi}(\tilde{\beta}_m, \tilde{v}_{n+1})$ are the stored values of Φ corresponding to the two adjacent $\tilde{v}$ values. The value of Φ corresponding to the given v_k can be approximated by following the linear-average equation $\Phi(\tilde{\beta}_m, v_k) \simeq \frac{1}{\tilde{v}_{n+1} - \tilde{v}_n} \{ (\tilde{v}_{n+1} - v_k) \tilde{\Phi}(\tilde{\beta}_m, \tilde{v}_n) + (v_k - \tilde{v}_n) \tilde{\Phi}(\tilde{\beta}_m, \tilde{v}_{n+1}) \}$. In addition, the value of β is set to the range of $[0.001, 4]$, with a step-size of 0.03125 to approximate a given β value. Thus we have 129 different β values reflected in the table of $\Phi(\beta, v_k)$. Consequently, the lookup table requires a memory space of a total of $129 \times 521 \times 32$ bits (floating point format), which is equivalent to about 268.8 kbytes.

3.8 Performance Evaluation

In Sections 3.3, 3.4 and 3.5, three adaptation methods based on our β -order MMSE estimation have been proposed. They are the adaptive β -order method, the adaptive β -order MMSE with speech-presence uncertainty, and the masking-based adaptive β -order MMSE method. To evaluate the performances of the proposed β -order MMSE speech enhancement methods, five different types of noise taken from the NOISEX-92 database are used in our simulation experiments. They are white Gaussian noise, interior Volvo car noise, F16 cockpit noise, Babble noise (100 people speaking in a canteen) and Leopard (military vehicle) noise. A total of 30 phonetically balanced speech utterances from the TIMIT database are used in our evaluation. Fifteen of the utterances are produced by male speakers and another fifteen by female speakers. The test utterances are sampled at 8 kHz and the frame size used is 256 samples (32 ms) in length. The samples are Hamming windowed with 50% overlap between adjacent frames. In order to simplify the noise estimation, we adopt minimum statistics [91]. We use α_{over} as the noise compensation factor, and the ratio (α_{over}/α_e) is set in the range of $1.0 \sim 1.5$ for different speech estimators.

As it is well known that segmental SNR (seg.SNR) [95] is more accurate in indicating the

²This range of v_k is sufficient to ensure that ξ_k ranges from about -15 to $+\infty$ dB, and instantaneous SNR ($\gamma_k - 1$) is about $-\infty$ to 35 dB.

speech distortion than the global SNR, noise was therefore added at different input seg.SNRs. In our simulation, each of the noisy utterances has an input seg.SNR. We have to set a group of utterances at a specified input seg.SNR level (e.g., -10 dB). This is done by computing an initial seg.SNR, SNR_I (in dB), with the corresponding initial multiplication factor P_I , where $x_I(t) = s(t) + P_I n(t)$. The noisy speech utterance at a desired input seg.SNR level can then be obtained by $x_d(t) = s(t) + P_d n(t)$, where $P_d = P_I 10^{(SNR_I - SNR_d)/20}$.

The speech enhancement algorithms for evaluation comparisons include Wiener filtering, over-subtraction masking [80], E-M LSA [118], E-M MMSE [115], optimal modified LSA (OM-LSA) [39], the adaptive β -order method [8] proposed in Section 3.3, the time-frequency adaptive β -order MMSE with/without speech-presence uncertainty proposed in Section 3.4, and the masking-based adaptive β -order MMSE method proposed in Section 3.5. Fig. 3.9 shows the spectrograms of the original clean speech, the input noisy speech and the enhanced speech obtained by the adaptive β -order method of Section 3.3, the masking-based adaptive β -order MMSE method of Section 3.5 as well as the various existing enhancement methods. Fig. 3.10 gives another example for the spectrograms of a different clean speech, the input noisy speech and the enhanced speech obtained by the time-frequency adaptive β -order method of Section 3.4 with/without speech-presence uncertainty, as well as other existing enhancement methods.

Figs. 3.11 and 3.12 show the seg.SNR [95], the average PESQ and the Itakura-Saito (IS) distortion measure ³ [95], respectively, of the enhanced speech processed by the different enhancement methods, for different input seg.SNRs (-10 dB, -5 dB, 0 dB, 5 dB and 10 dB) with white noise as well as F16 cockpit noise respectively. Figs. 3.13 and 3.14 as well as 3.15 and 3.16 show the PESQ performance, instead of improvement in average se.SNR, of the estimators with the same conditions used in 3.11 and 3.12.

From the obtained results, it is apparent that our proposed method almost always has the best seg.SNR, PESQ and IS distortion measure as compared to other methods.

It is well accepted that meaningful speech quality assessment must be based on human perception. Speech signals can be compared and evaluated in terms of intelligibility, naturalness,

³The highest 15% of the IS distortion values were discarded, to exclude the unrealistically high distortion values.

Figure 3.9: Speech spectrograms (male - "The fifth jar contains big, juicy peaches", 8 kHz sampling rate): (a) Clean Speech; (b) Noisy speech (F16 noise); (c) Wiener filtering method; (d) Over-subtraction masking method; (e) E-M LSA; (f) E-M MMSE; (g) Adaptive β -order MMSE; (h) Masking-based adaptive β -order MMSE.

Figure 3.10: Speech spectrograms (female - “The industry bargains as a unit with a single union”, 8kHz sampling rate): (a) Clean speech; (b) Noisy speech (F16 noise); (c) E-M LSA; (d) OM-LSA; (e) Adaptive β -order MMSE; (f) Time-frequency adaptive β -order MMSE without speech-presence uncertainty; (g) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (h) Masking-based adaptive β -order MMSE.

Figure 3.11: Improvement in seg.SNR for white noise. The tested methods are as following: (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; (□) Optimal modified LSA; (◇) E-M MMSE; (+) Adaptive β -order MMSE; (Δ) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (*) Proposed masking-based adaptive β -order MMSE.

Figure 3.12: Improvement in seg.SNR for F16 cockpit noise. The tested methods are as following: (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; (□) Optimal modified LSA; (◊) E-M MMSE; (+) Adaptive β -order MMSE; (Δ) Time-frequency adaptive β -order MMSE with speech-presence uncertainty; (*) Proposed masking-based adaptive β -order MMSE.

Figure 3.13: PESQ measure for white cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.

Figure 3.14: PESQ measure for F16 cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; ($\diamond$) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.

Figure 3.15: Itakura-Saito distortion measure for white noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; (◇) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.

Figure 3.16: Itakura-Saito distortion measure for F16 cockpit noise. The tested methods are as following: (.) Input noisy speech; (-.) Wiener filtering; (x) Over-subtraction masking; (o) E-M LSA; (◇) E-M MMSE; (+) Adaptive β -order MMSE; (*) Proposed masking-based adaptive β -order MMSE.

Table 3.1: MOS values for the following methods. 1) WF: Wiener filtering; 2) LSA: E-M Log spectral amplitude; 3) MMSE: E-M MMSE; 4) OM: Over-subtraction based on masking properties; 5) β : Proposed adaptive β -order MMSE; 6) U- β : Proposed time-frequency adaptive β -order MMSE with speech presence uncertainty; 7) M- β : Proposed masking-based adaptive β -order MMSE. The input seg.SNR is -5 dB with MOS = 1.85.

<i>Methods</i>	WF	LSA	MMSE	OM	β	U- β	M- β
MOS	2.01	3.25	2.98	2.81	3.39	3.38	3.52

and suitability for a particular application. The Mean Opinion Score (MOS) is one of the most popular and widely accepted subjective measures of “acceptability” that is influenced by the intelligibility and quality of the utterances being evaluated. Table 4.9 gives the MOS listening test results which confirm that our proposed masking-based enhancement method has the best performance for human listeners as compared to other enhancement methods.

To implement the adaptive β -order MMSE algorithm for real-time realization, the computational complexity involved in Eq. (3.16) can be simplified [6]. On the other hand, from the view point of processing complexity, the additional step of computing the masking threshold does not add too demanding a cost because our algorithm is working in the spectral domain using FFT.

3.9 Summary

In this chapter, the spectral suppression approach for speech enhancement has been studied. The β -order STSA MMSE estimation method has been proposed. The relation between the value of β and the spectral amplitude suppression gain function of the method is analyzed. The estimators with a series of constant β values are investigated through a large number of simulation experiments. In addition, β -order MMSE estimation with the adaptation of β value based on frame SNR is applied. A time-frequency method which adapts the value of β according to both frame and local SNRs is also proposed. It is found to be effective for preserving weak spectral components of a speech signal corrupted by noise. Moreover, the perceptual properties of the human auditory system was incorporated into β -order MMSE. The proposed method leads to an improvement over the SNR-based adaptive β -order MMSE method. Finally the musical noise elimination of β -order MMSE is analyzed and a lookup-table-search method is described

to reduce the computational complexity.

The computer simulation results clearly show that the proposed adaptive β -order MMSE methods outperform many existing speech enhancement methods significantly and have a good trade-off between speech distortion and noise reduction, especially for the case of weak spectral components of speech signal corrupted by noise. This conclusion is supported by visual spectrogram inspection, seg.SNR measure, PESQ and IS measures as well as MOS listening tests.

Chapter 4

Kalman Filtering Method

4.1 Introduction

Kalman filtering has been shown to be an effective speech enhancement technique [59, 53, 99, 110, 122], in which a speech signal is usually modeled by an autoregressive (AR) process and represented in the state-space domain. Various speech enhancement methods based on Kalman filtering have been reported in the literature. While many of the methods differ only by the choice of the algorithm used to estimate model parameters and the choice of the models employed for the speech signal and the additive noise, some of them introduce various perceptual factors to improve the quality of the estimated speech. Paliwal and Basu [59] estimated the speech parameters from clean speech first, then used a delayed version of Kalman filter to estimate the speech signal from a given noisy version contaminated by white noise. In [53], it is demonstrated that Kalman filtering with white noise assumption is not very effective in a colored noise environment, where the Kalman filter with colored noise model achieves expectedly better performance. In [110], it is shown that Kalman filtering-based speech enhancement in the subband domain not only greatly reduces the computation complexity, but also achieves a better performance as compared to that in the full-band domain. Taking into consideration implementation and complexity of a speech enhancement algorithm for possible applications such as mobile communications in a vehicle, Kalman filtering could be a good technique that deserves further research. The advantage of Kalman filtering, as compared to spectral suppression [98, 115], is that it can completely overcome the tonal noise problem and achieve quite good speech quality by reducing

the processing distortion introduced to a speech signal. Kalman filtering not only uses the statistical characteristics of signal and noise, but also utilizes the well-known autoregressive (AR) speech model which has been proven to be very effective for modelling the human speech production system [64].

In addition to Kalman filtering, the masking properties of the human auditory system have been applied successfully in spectral suppression methods [80, 52] and subspace techniques [21]. The main aim of these methods is to find an optimal trade-off between speech distortion and noise reduction. In spectral domain speech enhancement, the critical band concept and the masking threshold can be applied directly. In subspace-based speech enhancement, the masking threshold could be exploited via the relation between the eigen parameters and the power spectral density. This relation could be mathematically established according to Blackman-Tuckey cross-spectral analysis with a Bartlett window [21]. In [78], a Kalman filtering speech enhancement approach incorporating masking threshold is introduced. The time-domain Kalman filtering method is extended by calculating the noise masking threshold through Fourier transform, followed by inverse transform with re-estimation of speech signal based on the masking threshold. Therein, the masking threshold is used to reduce the inaudible noise by discarding the spectral components below it. In [79], the enhanced speech from a Kalman filter is further processed by a masking-based speech enhancement system implemented in the frequency domain. However, it does not help to preserve weak spectral components which might already have been removed by the preceding Kalman filter. Historically, noise masking in speech processing has quite a wide usage. In speech coding, it is used to shape quantization noise while maintaining the perceptual quality of the decoded speech. In speech enhancement, the objective is to reduce the noise that is perceptible while keeping the masked (or inaudible) noise unsuppressed in order to reduce speech distortion. This is because the main aim of speech enhancement is to maximize perceptual noise reduction at the cost of a certain level of speech distortion, or equivalently to minimize speech distortion while achieving a certain level of perceptual noise reduction. In some applications such as speech recognition which involves speech enhancement as a preprocessing step, speech

distortion is costlier than noise reduction.

In this chapter, a subband Kalman filtering scheme which incorporates the auditory masking properties for single channel speech enhancement is examined. It attempts to achieve high quality enhanced speech by optimizing the trade-off between speech distortion and noise reduction. The use of Kalman filtering in the subband instead of the full-band domain leads to considerable complexity reduction and performance improvement. A novel approach to incorporate the masking threshold within Kalman filtering is proposed, whereby the noise variance estimate used in the filtering process is modified according to the masking threshold in each subband. Through simulations, the proposed scheme is investigated by studying the contributions from different functions including Kalman filtering, subband decomposition and perceptual effect based on masking threshold. At the same time, the optimal configuration for the proposed scheme is examined. The proposed approach leads to better enhancement results as compared to the full-band and the normal subband Kalman filtering methods. The rest of this chapter is organized as follows. In Section 4.2, Kalman filtering for speech enhancement is reviewed. A subband Kalman filtering speech enhancement scheme based on the perceptual properties of the human auditory system is proposed in Section 4.3. In Section 4.4, the system configuration of the masking-based subband Kalman filtering method is investigated. Simulation results are given in Section 4.5, and the summary is provided in Section 4.6.

4.2 Kalman Filtering for Speech Enhancement

4.2.1 White Noise Model for Speech Enhancement

Consider the following model for noisy speech

$$x(n) = s(n) + v(n), \quad n = 1, 2, \dots \quad (4.1)$$

where $x(n)$ and $s(n)$ denote the discrete time samples of noisy speech and clean speech respectively, and $v(n)$ is a white Gaussian noise with variance σ_v^2 .

The clean speech signal is commonly modelled as an AR process, and approximated as the output of an all-pole linear system driven by an excitation signal, $w(n)$, which is assumed to be

a zero-mean white Gaussian process with variance σ_w^2 , i.e.,

$$s(n) = \sum_{i=1}^p a_i s(n-i) + w(n). \quad (4.2)$$

Let $\mathbf{s}(n) = [s(n-p+1) \ \dots \ s(n-1) \ s(n)]^T$, and AR parameter $\mathbf{a}(n) = [a_p \ \dots \ a_2 \ a_1]^T$.

We reformulate Eqs. (4.1) and (4.2) in the form of Kalman process and measurement equations in state space domain as follows:

$$\mathbf{s}(n) = \mathbf{F}\mathbf{s}(n-1) + \mathbf{g}w(n) \quad (4.3)$$

$$x(n) = \mathbf{C}^T \mathbf{s}(n) + v(n) \quad (4.4)$$

where

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ a_p & a_{p-1} & \dots & a_1 \end{bmatrix}_{p \times p}, \quad \mathbf{g} = \mathbf{C} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{p \times 1}. \quad (4.5)$$

Assuming that the parameters $\mathbf{a}(n) = [a_p \ \dots \ a_2 \ a_1]^T$, σ_w^2 and σ_v^2 are known, the optimal estimate of the state vector $\mathbf{s}(n)$ can then be obtained by Kalman filtering [59, 96] as follows:

$$\hat{\mathbf{s}}(n|\mathbf{x}_n) = \hat{\mathbf{s}}(n|\mathbf{x}_{n-1}) + \mathbf{G}(n)[x(n) - \mathbf{C}^T \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})] \quad (4.6)$$

$$\hat{\mathbf{s}}(n|\mathbf{x}_{n-1}) = \mathbf{F}\hat{\mathbf{s}}(n-1|\mathbf{x}_{n-1}) \quad (4.7)$$

$$\mathbf{K}(n) = [\mathbf{I} - \mathbf{G}(n)\mathbf{C}^T]\mathbf{K}(n, n-1) \quad (4.8)$$

where

$$\mathbf{G}(n) = \mathbf{K}(n, n-1)\mathbf{C}[\mathbf{C}^T\mathbf{K}(n, n-1)\mathbf{C} + \sigma_v^2]^{-1} \quad (4.9)$$

$$\mathbf{K}(n, n-1) = \mathbf{F}\mathbf{K}(n-1)\mathbf{F}^T + \mathbf{Q}. \quad (4.10)$$

The estimate, $\hat{s}(n)$, of the speech signal at time instant n is obtained by,

$$\hat{s}(n) = \mathbf{C}^T \hat{\mathbf{s}}(n|\mathbf{x}_n) \quad (4.11)$$

where $\hat{\mathbf{s}}(n|\mathbf{x}_{n-1})$ is the one-step prediction of the state vector $\mathbf{s}(n)$ at time n giving all past observations available up to and including time $n-1$, $\hat{\mathbf{s}}(n|\mathbf{x}_n)$ is the estimate of the state vector $\mathbf{s}(n)$ at time n , $\mathbf{G}(n)$ is the Kalman gain, and $\mathbf{K}(n, n-1) = \mathbf{E}\{[\mathbf{s}(n) - \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})][\mathbf{s}(n) - \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})]^T\}$ is the predicted state-error correlation matrix. $\mathbf{Q} = \mathbf{g}\mathbf{E}[w(n)w(n)]\mathbf{g}^T$ is a sparse matrix with only one nonzero element, that is $\mathbf{Q}(p, p) = \sigma_w^2$.

4.2.2 Colored Noise Model for Speech Enhancement

In order to model the colored noise, we presume that the noise is wide sense stationary and it is adequately modelled by using the AR process [53], i.e.,

$$v(n) = \sum_{i=1}^q b_i v(n-i) + u(n) \quad (4.12)$$

where $u(n)$ is a white Gaussian process with variance σ_u^2 . The AR parameters $\mathbf{b} = [b_q \ \dots \ b_2 \ b_1]^T$ and σ_u^2 could be estimated during a speech pause interval. Let $\mathbf{v}(n) = [v(n-q+1) \ \dots \ v(n-1) \ v(n)]^T$,

$$\mathbf{F}_v = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ b_q & b_{q-1} & \dots & b_1 \end{bmatrix}_{q \times q}, \quad \mathbf{g}_v = \mathbf{C}_v = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{q \times 1}. \quad (4.13)$$

Consequently, Eqs (4.1), (4.2) and (4.12) can be expressed in state-space domain as follows:

$$\bar{\mathbf{s}}(n) = \bar{\mathbf{F}}\bar{\mathbf{s}}(n-1) + \bar{\mathbf{g}}\bar{\mathbf{w}}(n) \quad (4.14)$$

$$x(n) = \bar{\mathbf{C}}^T \bar{\mathbf{s}}(n) \quad (4.15)$$

where

$$\bar{\mathbf{s}}(n) = \begin{bmatrix} \mathbf{s}(n) \\ \mathbf{v}(n) \end{bmatrix}, \quad \bar{\mathbf{w}}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix} \quad (4.16)$$

$$\bar{\mathbf{F}} = \begin{bmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_v \end{bmatrix}, \quad \bar{\mathbf{g}} = \begin{bmatrix} \mathbf{g} & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_v \end{bmatrix}, \quad \bar{\mathbf{C}} = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}_v \end{bmatrix}. \quad (4.17)$$

According to Kalman filtering theory, the estimation of state vector $\mathbf{s}(n)$ can be obtained by the following recursive equations:

$$\hat{\mathbf{s}}(n|\mathbf{x}_n) = \hat{\mathbf{s}}(n|\mathbf{x}_{n-1}) + \mathbf{G}(n)[x(n) - \bar{\mathbf{C}}^T \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})] \quad (4.18)$$

$$\hat{\mathbf{s}}(n|\mathbf{x}_{n-1}) = \bar{\mathbf{F}} \hat{\mathbf{s}}(n-1|\mathbf{x}_{n-1}) \quad (4.19)$$

$$\bar{\mathbf{K}}(n) = [\mathbf{I} - \mathbf{G}(n)\bar{\mathbf{C}}^T] \bar{\mathbf{K}}(n, n-1) \quad (4.20)$$

where

$$\mathbf{G}(n) = \bar{\mathbf{K}}(n, n-1) \bar{\mathbf{C}} [\bar{\mathbf{C}}^T \bar{\mathbf{K}}(n, n-1) \bar{\mathbf{C}}]^{-1} \quad (4.21)$$

$$\bar{\mathbf{K}}(n, n-1) = \bar{\mathbf{F}} \bar{\mathbf{K}}(n-1) \bar{\mathbf{F}}^T + \bar{\mathbf{Q}}. \quad (4.22)$$

The predicted state-error correlation matrix $\bar{\mathbf{K}}(n, n-1) = \mathbf{E}\{[\bar{\mathbf{s}}(n) - \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})][\bar{\mathbf{s}}(n) - \hat{\mathbf{s}}(n|\mathbf{x}_{n-1})]^T\}$. $\bar{\mathbf{Q}} = \bar{\mathbf{g}}\mathbf{E}[\bar{\mathbf{W}}(n)\bar{\mathbf{W}}(n)^T]\bar{\mathbf{g}}^T$ is a sparse matrix with only two nonzero elements; they are $\bar{\mathbf{Q}}(p, p) = \sigma_w^2$ and $\bar{\mathbf{Q}}(p+q, p+q) = \sigma_u^2$. The estimate of the speech signal is the output of Kalman filter, i.e.,

$$\hat{s}(n) = \mathbf{C}_1^T \hat{\mathbf{s}}(n|\mathbf{x}_n), \quad \mathbf{C}_1 = [\mathbf{C}^T \underbrace{0 \dots 0}_q]^T. \quad (4.23)$$

4.3 Incorporating Masking Properties into Subband Kalman Filtering

We propose a speech enhancement approach that incorporates the masking properties of the human auditory system into a subband Kalman filtering system. The noise masking threshold is mapped to each subband, and we use a Kalman filter for de-noising in the subband. The noise variance used for Kalman filtering in each subband is modified according to the mapped noise masking threshold in the subband, so as to control the degree of noise reduction of Kalman filtering.

4.3.1 Subband Filter

A uniform and critical polyphase modified discrete cosine transform (MDCT) filterbank method [105]-[86] is used for the design of the analysis and synthesis filter banks. The number of bands

is set to $\mathbb{B}$. The down sampling and up sampling rates are $M : 1$ and $1 : M$ respectively, and $M = \mathbb{B}$. If $\tilde{h}(n)$ represents the prototype filter of a subband, the τ -th polyphase filter $e_\tau(m)$ is then given by

$$e_\tau(m) = \tilde{h}(mM + \tau), \quad \tau = 0, 1, \dots, \mathbb{B} - 1. \quad (4.24)$$

The k -th analysis filter, $h_k(n)$, and k -th synthesis filter, $g_k(n)$, can then be proven to be given by

$$h_k(n) = h_k(mM + \tau) = 2e_\tau(m) \cos\left[\left(k - \frac{1}{2}\right)(\tau - 1)\frac{\pi}{M} + \phi_k(m) + (-1)^{k-1}\frac{\pi}{4}\right] \quad (4.25)$$

$$g_k(n) = g_k(mM + \tau) = 2e_\tau(m) \cos\left[\left(k - \frac{1}{2}\right)(\tau - 1)\frac{\pi}{M} + \phi_k(m) - (-1)^{k-1}\frac{\pi}{4}\right] \quad (4.26)$$

where $k = 0, 1, \dots, \mathbb{B} - 1$ and $\phi_k(m) = m\pi(k - \frac{1}{2})$.

In [105], it is shown that pseudo-QMF (quadrature mirror filter) bank with the stopband attenuation of the analysis (and thus synthesis) filters on the order of -100 dB can be achieved. Moreover, the resulting reconstruction error is also of the order of -100 dB. We use pseudo-QMF (quadrature mirror filter), which achieves near-perfect-reconstruction, to implement subband decomposition. The distortion and delay of the subband system depend on the selection of the prototype filter. We obtain the 4-band, 8-band, 16-band and 32-band prototype filter coefficients from the ISO 32-band Audio Band [41] prototype. Fig. 4.1 (a) shows the frequency response of the 16-band prototype filter. Figs. 4.1 (b) and (c) show the frequency responses of the 16-band analysis filters and their corresponding synthesis filters respectively, and Fig. 4.1 (d) shows the overall frequency distortion of this subband system.

4.3.2 AR Parameters Estimation in a Subband

In our study, we exploit subband Kalman filtering for speech enhancement, i.e., we apply Kalman filtering in each subband. As a result, we need to estimate the AR parameters for each subband. In particular, for the n -th sample of the k -th subband, the time-varying AR parameters $\mathbf{a} = [a_p \ \dots \ a_2 \ a_1]^T$, σ_w^2 and σ_v^2 (or $\mathbf{b} = [b_q \ \dots \ b_2 \ b_1]^T$ and σ_u^2) have to be estimated. As the AR parameters are critical for the accuracy of the Kalman filter's output, we adopt an iterative scheme to estimate the AR parameters. The initial AR parameters are estimated by using the

Figure 4.1: (a) The frequency response of the prototype filter for the 16-band configuration; (b) The frequency response of the 16-band analysis filters; (c) The frequency response of the 16-band synthesis filters; (d) The frequency response of the overall 16-band system.

noisy speech signal, and we refer to the first estimation as zeroth iteration. Subsequently, the AR parameters in the I -th iteration can be estimated by using the output signal of the $(I - 1)$ -th iteration of Kalman filtering.

The speech AR parameters $\mathbf{a} = [a_p \ \dots \ a_2 \ a_1]^T$ and colored noise AR parameters $\mathbf{b} = [b_q \ \dots \ b_2 \ b_1]^T$ can be obtained using the well known Levinson-Durbin algorithm [96].

In addition, the variance σ_w^2 could be estimated by the following approach:

$$\begin{aligned} \sigma_w^2 &= \mathbf{E}\left\{|s(n) - \sum_{i=1}^p a_i^*(n)s(n-i)|^2\right\} \\ &= R_{ss}(0) - 2\text{Re}\left\{\sum_{i=1}^p a_i^* R_{ss}(i)\right\} + \sum_{i=1}^p \sum_{j=1}^p a_i^* a_j R_{ss}(i-j). \end{aligned} \quad (4.27)$$

The AR parameters $\mathbf{a}$ and the autocorrelation values, R_{ss} , are approximated as follows:

$$\check{S}_{PSD}(\bar{f}_s, k, l) = |\check{S}(\bar{f}_s, k, l)|^2 \quad (4.28)$$

$$\hat{R}_{ss}(i) = \text{IDFT}\{\check{S}_{PSD}(\bar{f}_s, k, l) | \bar{f}_s = 1 : N_{sub}\}, \quad i = 0, \dots, p \quad (4.29)$$

where l is the frame number that covers the n -th sample of a subband, k denotes the subband

number, N_{sub} denotes the total number of frequency bins of the subband, and $\bar{f}_s$ denotes the frequency bin of the subband; $\check{S}_{PSD}(\bar{f}_s, k, l)$ and $\check{S}(\bar{f}_s, k, l)$ are the power spectral densities (PSD) and Fourier Transform of the estimated speech signal of the previous iteration, respectively. For the zeroth iteration, they represent the PSD and Fourier transform of the noisy signal, $x(n)$, respectively. $\hat{R}_{ss}$ is the estimate of the autocorrelation function of $s(n)$. The linear prediction coefficient (LPC) vector, $\mathbf{a}$, satisfy the Yule-Walker equations [24, 25] which are given by

$$\begin{bmatrix} R_{ss}(0) & R_{ss}(1) & \cdots & R_{ss}(p-1) \\ R_{ss}^*(1) & R_{ss}(0) & \cdots & R_{ss}(p-2) \\ \vdots & \vdots & \cdots & \vdots \\ R_{ss}^*(p-1) & R_{ss}^*(p-2) & \cdots & R_{ss}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_{ss}(1) \\ R_{ss}(2) \\ \vdots \\ R_{ss}(p) \end{bmatrix}. \quad (4.30)$$

The estimate of the LPC vector, $\mathbf{a}$, is computed by using the Levinson-Durbin recursive algorithm [54, 77, 96].

The PSD of noise signal is estimated as follows. Let the pre-SNR, $\rho_{pre}(\bar{f}_s, k, l)$, be

$$\rho_{pre}(\bar{f}_s, k, l) = 10 \log_{10} \max \frac{\{\max[|X(\bar{f}_s, k, l)| - D(\bar{f}_s, k, l-1), \quad \varepsilon]\}^2}{D(\bar{f}_s, k, l-1)^2} \quad (4.31)$$

where ε is a small value that is used to avoid negative infinity of logarithm when a zero occurs, and $D(\bar{f}_s, k, l)$ is the estimate of noise spectral amplitude that is approximately by the following smoothed periodogram scheme:

$$D(\bar{f}_s, k, l) = \begin{cases} \alpha_n D(\bar{f}_s, k, l-1) + (1 - \alpha_n) |X(\bar{f}_s, k, l)|, & \text{when } \rho_{pre}(\bar{f}_s, k, l) < \Omega \\ D(\bar{f}_s, k, l-1), & \text{when } \rho_{pre}(\bar{f}_s, k, l) \geq \Omega \end{cases} \quad (4.32)$$

where $X(\bar{f}_s, k, l)$ is the spectrum component of the noisy speech at bin $\bar{f}_s$, parameter α_n denotes the smoothing factor which is a forgetting factor in the recursive equation, and the threshold Ω is used to identify the absence period of a speech signal at a certain frequency in a subband. In our simulation, $\alpha_n = 0.89$ and $\Omega = 0$ dB.

For the colored noise model, the LPC parameters for the noise signal, $\mathbf{b}$, are only estimated in the zeroth iteration, and they will be kept unchanged in the subsequent iterations. The procedure is similar to that for $\hat{\mathbf{a}}$, i.e.,

$$\check{N}_{PSD}^{(0)}(\bar{f}_s, k, l) = |D^{(0)}(\bar{f}_s, k, l)|^2 \quad (4.33)$$

$$\hat{R}_N^{(0)} = \text{IDFT}\{\tilde{N}_{PSD}^{(0)}(\bar{f}_s, k, l) | \bar{f}_s = 1 : N_{sub}\} \quad (4.34)$$

$$\hat{\mathbf{b}} = \begin{bmatrix} \hat{R}_N^{(0)}(0) & \hat{R}_N^{(0)}(1) & \cdots & \hat{R}_N^{(0)}(q-1) \\ \hat{R}_N^{(0)}(1) & \hat{R}_N^{(0)}(0) & \cdots & \hat{R}_N^{(0)}(q-2) \\ \vdots & \vdots & \cdots & \vdots \\ \hat{R}_N^{(0)}(q-1) & \hat{R}_N^{(0)}(q-2) & \cdots & \hat{R}_N^{(0)}(0) \end{bmatrix}^{-1} \begin{bmatrix} \hat{R}_N^{(0)}(1) \\ \hat{R}_N^{(0)}(2) \\ \vdots \\ \hat{R}_N^{(0)}(q) \end{bmatrix} \quad (4.35)$$

where the superscript $^{(0)}$ represents the zeroth iteration in the estimate of the parameters. Fig. 4.2 shows the flow chart for the AR parameters estimation (LPC and excitation variance(s)) of the k -th subband for the I -th iteration.

It has been demonstrated that Kalman algorithms with white noise assumption are ineffective in colored noise environments [53]. The results of [53] also indicate that Kalman filters with a colored noise model provide a significant improvement in output speech quality. Therefore, for the purpose of performance comparison, we use Kalman filter with a white noise model for white noise cases and apply Kalman filter with a colored noise model for colored noise cases.

The order p of the AR model is determined by the number of formants of a speech phoneme. For an 8 kHz speech signal, p value for full-band is about 10. The AR order will determine the complexity of the enhancement algorithm and the processed speech quality of a Kalman filtering speech enhancement system. Actually, the AR order, p , determines how accurately an AR linear prediction model describes the spectral envelope of a signal. Fig. 4.3 gives an example of speech spectral magnitude and the spectral envelop with $p = 5$ and 16 respectively.

4.3.3 Incorporating Masking into Subband Kalman Filtering

The noise masking threshold can be obtained through modelling the frequency selectivity of the human auditory system and its masking properties. The procedure of computing a noise masking threshold, $\Psi(C_{bark})$, involves Bark (or critical) band analysis, a spreading function, the relative threshold offset and the absolute auditory threshold, as described in Section 2.3.

After the noise masking threshold is obtained, the normalized masking threshold, $\bar{\Psi}$, is given as follows

$$\bar{\Psi}(C_{bark}, \tilde{l}) = \frac{\Psi(C_{bark}, \tilde{l}) - \Psi_{min}(\tilde{l})}{\Psi_{max}(\tilde{l}) - \Psi_{min}(\tilde{l})} \quad (4.36)$$

Figure 4.2: The flow chart for speech AR parameters estimation at the I -th iteration in the k -th subband.

where $\Psi_{max}(\tilde{l})$ and $\Psi_{min}(\tilde{l})$ are respectively the maximum and minimum values of the noise masking threshold in the full-band frame $\tilde{l}$. The normalized masking threshold in the Bark domain has its corresponding expression $\bar{\Psi}_f$ in the frequency domain, i.e.,

$$\bar{\Psi}_f(\bar{f}, \tilde{l}) = \bar{\Psi}(C_{bark}, \tilde{l}) \quad (4.37)$$

where $\bar{f}$ denotes the full-band frequency bin. Just like most masking-based speech enhancement methods, the noise masking threshold $\bar{\Psi}(C_{bark}, \tilde{l})$ has to be obtained from a speech signal. In

Figure 4.3: An example of the power spectral magnitude of a speech signal (32 ms), and its spectral envelopes obtained by using AR parameters with $p=5$ and $p=16$ respectively.

the zeroth iteration, $\Psi(C_{bark}, \tilde{l})$ is obtained from the noisy speech signal, and in the subsequent iterations it is obtained from the estimated speech signal of the previous iteration.

The relation between the uniform subband frequency bin, $\bar{f}_s$, and the full-band frequency bin, $\bar{f}$, is given by

$$\bar{f} = \frac{N_{full}}{\mathbb{B}}(k + \frac{\bar{f}_s}{N_{sub}}) \quad (4.38)$$

where N_{full} indicates the total number of frequency bins of the full-band¹. The masking threshold, $\Psi_s(\bar{f}_s, k)$, corresponding to frequency $\bar{f}_s$ in the k -th subband can therefore be obtained by

$$\bar{\Psi}_s(\bar{f}_s, k) = \bar{\Psi}_f(\frac{N_{full}}{\mathbb{B}}(k + \frac{\bar{f}_s}{N_{sub}}), \tilde{l}). \quad (4.39)$$

Though most speech enhancement algorithms succeed to a great extent in reducing audible noise, they also attenuate masked noise and the original speech signal. Consequently, the perceptual effect of the distorted speech may be more noticeable than that of the reduced noise. Assume that $G_{\bar{f}_s}$ denotes the suppression gain at frequency bin $\bar{f}_s$. $S_{\bar{f}_s}$ is the $\bar{f}_s$ th spectral component of clean speech signal and $\hat{S}_{\bar{f}_s}$ is an estimate of $S_{\bar{f}_s}$. The distortion is $\hat{S}_{\bar{f}_s} - S_{\bar{f}_s} = G_{\bar{f}_s}(S_{\bar{f}_s} + N_{\bar{f}_s}) - S_{\bar{f}_s} = S_{\bar{f}_s}(G_{\bar{f}_s} - 1) + G_{\bar{f}_s}N_{\bar{f}_s} = r_s(\bar{f}_s) + r_n(\bar{f}_s)$, where $r_s(\bar{f}_s) \triangleq S_{\bar{f}_s}(G_{\bar{f}_s} - 1)$ represents speech distortion, $r_n(\bar{f}_s) \triangleq G_{\bar{f}_s}N_{\bar{f}_s}$ is the residual noise, and gain $G_{\bar{f}_s}$ is in the range of $[0 \ 1]$. If we do not perform suppression, i.e., $G_{\bar{f}_s} = 1$, the speech distortion part $r_s(\bar{f}_s)$ is 0, but at the cost of unsuppressed noise $r_n(\bar{f}_s) = N_{\bar{f}_s}$. When we suppress more noise (i.e., smaller $G_{\bar{f}_s}$), we get additional speech distortion. The critical issue therefore lies on how to select an optimal trade-off between speech distortion and noise reduction. A number of papers [15, 80] have addressed this issue and the main approach considered involves the use of masking properties. The basic idea is to use the auditory masking threshold to regulate the suppression gain for a frequency $\bar{f}_s$. When the masking threshold is high, the gain value is raised. When it is low, the gain value is reduced.

In our proposed Kalman filtering speech enhancement method, the regulation of noise reduction for speech enhancement is tied to the level of masking threshold. Through a large number of

¹Obviously, the corresponding relation between the linear frequency f and the frequency bin $\bar{f}$ is $\bar{f} = \frac{f * N_{full}}{\text{sampling rate}/2}$.

simulations, it is found that over-estimation of noise variance would lead to more suppression of noise, and under-estimation of the noise variance would result in less attenuation of noise. Therefore the regulation of the noise variance leads to the control of gain indirectly. When the masking threshold is high, we under-estimate the noise to decrease the degree of noise reduction; when the masking threshold is low, we reduce the degree of under-estimation or even over-estimate the noise. The control of the noise variance value can be achieved by multiplying it by a masking threshold dependent factor, which has been found to be effective in controlling the trade-off between speech distortion and noise reduction. From the above analysis, we propose that the PSD of the noise signal be modified as a linear function of masking threshold, i.e.,

$$N_{PSD}(\bar{f}_s, k, l) = \{\alpha_{t_1}[\alpha_{t_2} - \bar{\Psi}_s(\bar{f}_s, k)]D^{(0)}(\bar{f}_s, k, l)\}^2 \quad (4.40)$$

where α_{t_1} and α_{t_2} are constants which regulate the relation between noise PSD and masking threshold. α_{t_1} and α_{t_2} are chosen so that when the masking threshold at the frequency bin $\bar{f}_s$ reaches the minimum value among the full-band masking threshold, the modified PSD, $N_{PSD}(\bar{f}_s, k, l)$, should be equal to its original value, $D^{(0)}(\bar{f}_s, k, l)\}^2$. In this case, we have $\alpha_{t_1} = \frac{1}{\alpha_{t_2}}$. On the other hand, the normalized masking threshold, $\bar{\Psi}_s(\bar{f}_s, k)$, is in the range of $[0, 1]$. When the threshold value reaches its maximum (i.e., $\bar{\Psi}_s(\bar{f}_s, k)=1$), the PSD should be modified the most with the use of the minimum multiplication factor δ_{min} , i.e., $\alpha_{t_1}[\alpha_{t_2} - 1] = \delta_{min}$. Empirically δ_{min} is set to $\frac{2}{3}$. Accordingly, α_{t_1} and α_{t_2} could have been set to $\frac{1}{3}$ and 3 respectively. However, according to our simulations, it is found that when the masking threshold is low, a small over-estimation will give rise to a better performance of speech enhancement. We therefore finally set α_{t_1} and α_{t_2} to be 0.35 and 3.0 respectively. $N_{PSD}(\bar{f}_s, k, l)$ is estimated in each iteration as $\bar{\Psi}_s(\bar{f}_s, k)$ is updated in each iteration, and $D^{(0)}(\bar{f}_s, k, l)$ is only estimated during the zeroth iteration, i.e., when the original noisy speech signal is used as the reference input.

Subsequently, the modified noise autocorrelation can be obtained by

$$R_N(i) = \text{IDFT}\{N_{PSD}(\bar{f}_s = 1 : N_{sub}, k, l)\}, \quad i = 0, 1, \dots, q. \quad (4.41)$$

For the white noise model, the modified noise variance, σ_v^2 , is given by

$$\sigma_v^2 = R_N(0) \quad (4.42)$$

and for the colored noise model, the modified noise variance, σ_u^2 , is given by

$$\begin{aligned} \sigma_u^2 &= \mathbf{E}\left\{|v(n) - \sum_{i=1}^q b_i^*(n)v(n-i)|^2\right\} \\ &= R_N(0) - 2\text{Re}\left\{\sum_{i=1}^q b_i^* R_N(i)\right\} + \sum_{i=1}^q \sum_{j=1}^q b_i^* b_j R_N(i-j). \end{aligned} \quad (4.43)$$

Fig. 4.4 gives the flow chart which shows the iterative computation of the modified noise variance in a subband.

Figure 4.4: The flow chart of the modified noise variance (or AR parameters for colored noise) estimation at the I -th iteration in the k -th subband. (a) white noise model; (b) colored noise model.

In exploiting the simultaneous frequency masking properties, the common approach involves the determination of masking threshold for each critical band. This implies the need for non-uniform subband decomposition, which leads to high computation cost. A simpler alternative

is to perform uniform subband decomposition and map approximately the masking threshold in each critical band to the corresponding subbands. Obviously, the higher the number of uniform subbands, the more accurate would be the mapping between the uniform band-edges and the critical band-edges. In this study, we select 8, 16 and 32 uniform subbands with 8 kHz sampling rate in our simulations. The block diagram of our masking-based speech enhancement method is shown in Fig. 4.5.

In order to track the progress of the AR parameters estimation at each iteration, we show the spectral envelope of the enhanced speech in each iteration of the proposed masking-based subband Kalman filtering method with the colored noise model ($\mathbb{B} = 16$ and $(p, q) = (1, 1)$) for the F16 (colored) noise case in Figs. 4.6 and 4.7 for both a vowel and a consonant respectively.

Figure 4.5: Structure of the proposed masking-based subband Kalman filtering scheme ('pre-IT' means the previous iteration, and IT denotes the number of iterations), where the double-lines represent the path of feedback from the previous iteration to the current iteration.

Figure 4.6: Vowel 'i' (8kHz sampling rate), with the input noisy SNR = 3.5 dB. (a) waveform of the vowel 'i', (b) ideal case AR spectral envelope, (c) zeroth iteration AR spectral envelope, (d) first iteration AR spectral envelope, (e) 3rd iteration AR spectral envelope, (f) 5th iteration AR spectral envelope.

Figure 4.7: consonant 's' (8kHz sampling rate), with the input noisy SNR = -10 dB. (a) waveform of the consonant 's', (b) ideal case AR spectral envelope, (c) zeroth iteration AR spectral envelop, (d) first iteration AR spectral envelop, (e) 3rd iteration AR spectral envelop, (f) 5th iteration AR spectral envelop.

Table 4.1: IMPROVEMENT IN SEGMENTAL SNR OF ENHANCED SPEECH, FOR COMPARISON OF DIFFERENT BAND NUMBER, WITH NOISY INPUT SEGMENTAL SNR = -5 dB

$\mathbb{B}$	White Gaussian Noise $dB / (p)$		Colored (Volvo) Noise $dB / (p, q)$	
	zeroth-IT	ideal	zeroth-IT	ideal
2	5.89 / (8)	12.65 / (8)	6.65 / (8,5)	16.50 / (8,5)
4	6.09 / (5)	12.79 / (5)	7.15 / (5,3)	16.01 / (5,3)
8	6.93 / (3)	12.95 / (3)	8.81 / (3,2)	16.70 / (3,2)
16	6.58 / (1)	13.71 / (1)	8.73 / (1,1)	16.95 / (1,1)
32	6.29 / (1)	13.50 / (1)	8.53 / (1,1)	16.35 / (1,1)

4.4 System Configuration of Proposed Scheme

In our investigation of the system configuration, the frame size used for the full-band masking threshold calculation is 256 samples (32 ms) at 8 kHz sampling rate, and the frame size used for spectral estimation in each subband is $\frac{256}{\mathbb{B}}$ samples. In our simulation, we use a rectangular window for each frame in a subband. In order to determine the optimum system configuration for our proposed method, we use 6 speech utterances (3 male & 3 female) and two types of noise (white Gaussian noise and Volvo car noise (colored noise)) for performance analysis.

4.4.1 Determination of the Number of Subbands

First, we compare the effectiveness of the proposed system with different numbers of subbands in terms of average segmental SNR improvement. In this comparison, we use sufficiently large AR order for the different cases. Table 4.1 gives the results of our proposed method with different numbers of subbands for white and colored noise. From the results obtained, we can see the performance of the $\mathbb{B}=16$ ($p=1$) case is the best. Although the segmental SNR improvement with $\mathbb{B}=8$ ($p=3$) is higher than that with $\mathbb{B}=16$ ($p=1$) in the first estimation (zeroth iteration) for both white and colored noise cases, the former is lower than the latter in the ideal case. We thus chose $\mathbb{B}=16$ ($p=1$) masking-based Kalman filtering method for evaluation comparison.

4.4.2 Selection of AR Order

A Kalman filter can be used to predict or to estimate a signal. When it is used as a predictor, increasing AR order is always going to decrease the error signal, if the AR parameters are directly

obtained from the clean speech. In speech enhancement, Kalman filtering is exploited to filter out the noise and estimate the speech signal, and the AR parameters are only estimated from the noisy signal. In this case, effective estimation of the AR parameters requires the number of samples in a frame to be much larger than the order of AR, p . Considering the characteristics of 32 ms short-term stationary of a speech signal, we therefore have 16 samples for each of the 16 subbands for 8 kHz sampling rate. Since there are only 16 samples, an AR order of more than 2 or 3 could not be accurately determined. On the other hand, most of the subbands have no spectral peaks for the 16-subband case, even for voiced speech signals. For unvoiced speech signals, this situation is worse. In these cases, simply increasing the AR order is not helpful, since the spectrum is essentially flat or somewhat tilted, but not peaky. Similar observations pertaining to the AR order can be made for the colored noise case. Table 4.2 gives the results obtained with different AR orders for both white and colored noise cases. From the table, we can also see that the proposed method with $p=1$ and $\mathbb{B}=16$ seems to be well suited for the white noise model and the method with $(p, q) = (1, 1)$ and $\mathbb{B}=16$ seems to be the proper choice for the colored noise model. $(p, q) > (1, 1)$ leads only to a slightly better performance than $(p, q) = (1, 1)$, but with a higher computation cost. It is also interesting to note that the segmental SNR improvements with $p=1$ for the white noise case and $(p, q) = (1, 1)$ for the colored noise case are higher than those with $p > 1$ and $(p, q) > (1, 1)$ in the zeroth iteration, although the main concern is the final performance at the last iteration step.

4.4.3 Verification of Various Functions

Since our proposed method involves the use of Kalman filtering, subband decomposition and masking properties, we have carried out a number of simulations to show the individual effects of each of these functions separately. Tables 4.3 and 4.4 show the average performance results for the 6 speech utterances in terms of segmental SNR improvement and PESQ score. In the simulation, we choose $\mathbb{B}=16$ with AR order $p = 1$ for the white noise case and $(p, q) = (1, 1)$ for the colored noise case. The accuracy of the LPC coefficients of the AR parameters are found not to affect the final results strongly, provided that the driving noise variances are properly

estimated. On the other hand, both the subband and masking functions have been found to affect the final speech quality more significantly. As one of comparative counterparts, we choose the conventional full-band Kalman filtering [53] method with AR order $p = 16$ for white noise and $(p, q) = (16, 5)$ for colored noise ².

4.4.4 Monitoring of the Effect in Each Iteration

For the purpose of predicting the best and the worst performance of our proposed method, we first consider two extreme cases. One extreme case is when the time-varying parameters of the speech model are estimated using the clean speech instead of the noisy one, and we call it the “ideal” case. The other extreme case is when the AR parameters are estimated using the noisy speech without any iteration, which we identify as “zeroth-IT”, i.e., the zeroth iteration. For comparison, we also consider the same two extreme cases with the conventional Kalman filtering method. We further need to compare the various iterations of the enhanced speech obtained by using the conventional Kalman filtering method with that obtained by using the proposed masking-based subband Kalman filtering method, both with white and colored noise. In each iteration, we use the enhanced speech signal from the previous iteration to estimate the speech parameters, and still use the original noisy signal as the input signal.

We compare the performances of our proposed method and the full-band Kalman filtering method for different iterations as well as in the ideal case, where the AR parameters are obtained from the clean speech signals. The comparison is given in Tables 4.5 and 4.6 which show the average segmental SNR improvement and PESQ score for the enhanced speech obtained with both methods at the zeroth to fifth iteration as well as in the ideal case, using 6 speech utterances. Results given in both tables indicate that our proposed method is consistently superior to the conventional full-band Kalman filtering method. It can be observed that the best results are obtained with around five iterations (i.e., the fourth iteration) for the conventional full-band Kalman filtering method and three iterations (the second iteration) for the proposed masking-

²The $\mathbb{B}=16$ with AR order $p = 1$ of subband processing is equivalent to the $p = 16$ of full-band AR order. Although the conventional $p = 10$ AR order is enough for the 8 kHz sampling rate, in order to compare the effectiveness among those subband and full-band Kalman filtering methods under equivalent AR order conditions, we chose the $p = 16$ for full-band AR order.

based subband Kalman filtering method, and the segmental SNR improvement and PESQ score tend to drop thereafter.

4.5 Performance Evaluation

Performance evaluation of the proposed enhancement method is based on: i) visual spectrogram inspection; ii) objective assessment in terms of average segmental SNR improvement (obtained by computing the SNR for each speech frame and averaging the ratios over the entire signal [102]), and PESQ score; and iii) subjective assessment in terms of informal listening tests. Two different noise types, taken from the NOISEX-92 database [2], are used in our performance evaluation. They are white Gaussian noise, colored noise (interior Volvo car noise). Thirty phonetically balanced speech utterances from the TIMIT database [45] are used in our evaluation. Half of the utterances are spoken by male speakers and half by female. The effectiveness of the enhancement methods is evaluated at the sampling rate of 8kHz which is down-sampled from 16 kHz after pre-filtering.

Based on the simulation results obtained in the last Section, we now compare our proposed method (with $\mathbb{B}=16$, $p=1$ at the second iteration for white noise and $(p, q) = (1, 1)$ at the second iteration for colored noise) with other existing enhancement methods.

Considering the accuracy and reliability of the speech enhancement system, we use clean speech as an input, and compare the distortions introduced by our proposed method with those by other existing speech enhancement methods. It was found that the proposed masking-based subband Kalman filtering method has a very similar level of distortion as other existing enhancement methods such as spectral subtraction, E-M MMSE and full-band Kalman filtering in terms of segmental SNR, Itakura Saito measure and PESQ score.

For the conventional full-band Kalman filtering method, we still use its result obtained at its fourth iteration. Tables 4.7 and 4.8 compare in detail the results of different speech enhancement methods including E-M MMSE [115], over-subtraction masking [80], the conventional full-band Kalman filtering method [53], full-band Kalman post-masking [79]³, subband Kalman filtering

³In [79], mixed Kalman filtering, in which the measurement noise is colored noise instead of traditional white

[110] and our proposed masking-based subband Kalman filtering method. It is apparent that our proposed method always outperforms the other methods in terms of average segmental SNR improvement and PESQ score, for the 30 speech utterances taken from the TIMIT database.

We further make a comparison of spectrograms among various speech enhancement methods. Fig. 4.8 shows the spectrograms of clean speech, noisy speech which is degraded by white noise, speech enhanced by spectral subtraction [98], E-M MMSE [115], over-subtraction masking [80], the conventional full-band Kalman filtering method [53], subband Kalman filtering [110] and our proposed masking-based subband Kalman filtering method. Fig. 4.9 shows the spectrograms of the noisy speech which contains the same clean speech as in Fig. 4.8 and is degraded by colored (Volvo car) noise, and speech enhanced by the over-subtraction masking [80], the conventional full-band Kalman method [53], full-band Kalman post-masking [79], subband Kalman filtering method [110] and our proposed method.

Subjective listening tests that we have conducted also yield similar findings. In particular, undesirable musical noise can be heard in the enhanced speech obtained with spectral subtraction, but it completely disappears in the enhanced speech obtained by our proposed method. Moreover, the enhanced speech obtained with our proposed method demonstrates clarity and naturalness whereas that obtained with the conventional full-band Kalman filtering method sounds somewhat distorted, slightly slurred and muffled. The Mean Opinion Score (MOS) is one of the most popular and widely accepted subjective measures. In our subjective performance assessment, 40 listeners were invited to assess the subjective quality. Two speech utterances, with one from a female speaker and the other from a male speaker, contaminated by Volvo car noise at input seg.SNR of -5 dB are processed by different enhancement methods. Table 4.9 gives the MOS listening test results.

noise, is used only once and then followed by post filtering based on masking threshold. In order to compare fairly with conventional full-band Kalman methods, we use the fourth iteration of the full-band mixed Kalman filtering followed by post filtering based on masking properties. It is found that the post filtering step based on masking properties does not lead to improvement in PESQ after the fourth iteration. However, it does improve PESQ following the first iteration of Kalman filtering.

Table 4.2: IMPROVEMENT IN SEGMENTAL SNR OF ENHANCED SPEECH, FOR COMPARISON OF DIFFERENT AR ORDER (p, q) , WITH NOISY INPUT SEGMENTAL SNR = -5 dB, $\mathbb{B} = 16$

White Gaussian Noise dB (p)		Colored (Volvo) Noise dB (p, q)	
zeroth-IT	ideal	zeroth-IT	ideal
6.58 (1)	13.71 (1)	8.73 (1,1)	16.95 (1,1)
6.97 (2)	13.75 (2)	8.69 (2,1)	16.98 (2,1)
6.80 (3)	13.82 (3)	8.52 (3,1)	17.01 (3,1)
6.72 (5)	13.90 (5)	8.13 (5,2)	17.06 (5,2)
5.85 (9)	13.86 (9)	8.03 (9,5)	17.13 (9,5)

Table 4.3: SEGMENTAL SNR IMPROVEMENT AND PESQ SCORE OF ENHANCED SPEECH FOR WHITE GAUSSIAN NOISE, FOR COMPARISON OF DIFFERENT EFFECTS OF FUNCTIONS, NOISY INPUT SEGMENTAL SNR = -5 dB AND PESQ = 1.63. FK: FULL-BAND KALMAN FILTERING METHOD ($p=16$); SK: SUB-BAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, p=1$); MSK: THE PROPOSED MASKING-BASED SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, p=1$).

Method	Improvement in segmental SNR(dB)			PESQ score		
	zeroth-IT	IT=2	ideal	zeroth-IT	IT=2	ideal
FK	5.80	7.90	12.57	1.75	2.16	2.91
SK	6.12	10.30	13.08	1.95	2.43	2.93
MSK	6.58	11.16	13.71	2.01	2.63	3.05

Table 4.4: SEGMENTAL SNR IMPROVEMENT AND PESQ SCORE OF ENHANCED SPEECH FOR COLORED (VOLVO) NOISE, FOR COMPARISON OF DIFFERENT EFFECTS OF FUNCTIONS, NOISY INPUT SEGMENTAL SNR = -5 dB AND PESQ = 3.13. FK: FULL-BAND KALMAN FILTERING METHOD ($(p, q)=(16, 5)$); SK: SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, (p, q)=(1, 1)$); MSK: THE PROPOSED MASKING-BASED SUBBAND KALMAN FILTERING METHOD ($\mathbb{B} = 16, (p, q)=(1, 1)$).

Method	Improvement in segmental SNR(dB)			PESQ score		
	zeroth-IT	IT=2	ideal	zeroth-IT	IT=2	ideal
FK	6.62	12.20	16.43	3.20	3.38	3.93
SK	8.02	13.85	16.62	3.23	3.68	3.97
MSK	8.73	14.63	16.95	3.28	3.76	4.01

Table 4.5: Comparison between full-band Kalman filtering ($p=16$) and the proposed method ($\mathbb{B}=16, p=1$), with White Gaussian Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB and 5 dB

	IT no.	Improvement in segmental SNR(dB)				PESQ score			
Noisy speech		-10	-5	0	5	1.21	1.63	2.05	2.51
Enhanced speech, Fullband Kalman	zeroth-IT	6.80	5.80	4.71	3.73	1.36	1.79	2.33	2.73
	1	8.51	6.92	5.53	4.05	1.51	1.93	2.46	2.89
	2	10.26	7.90	6.35	4.33	1.67	2.16	2.55	2.95
	3	11.58	9.38	7.21	4.85	1.92	2.27	2.68	3.08
	4	11.93	9.96	7.58	4.98	1.93	2.29	2.65	3.06
	5	11.90	9.93	7.36	4.81	1.89	2.24	2.61	3.05
	ideal	15.38	12.57	9.68	6.91	2.52	2.91	3.25	3.65
Enhanced speech, our proposed method	zeroth-IT	7.83	6.58	5.16	3.98	1.60	2.01	2.52	2.91
	1	11.35	9.37	7.21	5.15	2.22	2.45	2.72	3.03
	2	13.69	11.16	8.62	6.65	2.39	2.63	2.98	3.27
	3	13.82	11.08	8.50	6.42	2.33	2.58	2.85	3.25
	4	13.73	10.98	8.38	6.39	2.30	2.46	2.80	3.19
	5	13.68	10.90	8.21	6.28	2.25	2.43	2.78	3.12
	ideal	16.53	13.71	10.63	7.59	2.68	3.05	3.38	3.70

Table 4.6: Comparison between full-band Kalman filtering ($(p, q)=(16, 5)$) and proposed method ($\mathbb{B}=16, (p, q)=(1, 1)$), with colored (Volvo) Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB and 5 dB

	Number of iteration	Improvement in segmental SNR(dB)				PESQ score			
Noisy speech		-10	-5	0	5	2.77	3.13	3.47	3.80
Enhanced speech, Full-band Kalman	0	7.92	6.62	5.63	4.25	2.81	3.20	3.56	3.91
	1	12.20	9.95	7.75	5.58	2.96	3.29	3.64	3.98
	2	15.67	12.20	9.28	6.12	3.05	3.38	3.73	4.03
	3	16.40	13.46	10.35	7.95	3.18	3.48	3.82	4.07
	4	16.78	13.68	10.76	8.16	3.22	3.60	3.90	4.08
	5	16.75	13.62	10.60	8.08	3.19	3.58	3.87	4.07
	ideal	18.32	16.43	13.78	10.69	3.55	3.93	4.18	4.28
Enhanced speech, our proposed method	0	10.70	8.73	7.32	5.45	2.85	3.28	3.61	3.95
	1	15.28	12.26	9.29	7.03	3.08	3.50	3.85	4.10
	2	17.15	14.63	11.82	9.33	3.31	3.76	4.02	4.17
	3	17.25	14.57	11.70	9.28	3.35	3.73	3.98	4.13
	4	17.12	14.40	11.58	9.13	3.25	3.67	3.93	4.10
	5	17.05	14.33	11.52	9.05	3.21	3.58	3.89	4.06
	ideal	18.93	16.95	14.32	11.28	3.65	4.01	4.26	4.32

Table 4.7: Segmental SNR improvement and PESQ of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($p=16$, $IT=4$); KM: Full-band Kalman post-masking ($(p=10)$, $IT=4$); SK: Subband Kalman ($p=1$, $\mathbb{B}=16$, $IT=2$); and MSK: Our proposed Method ($p=1$, $\mathbb{B}=16$, $IT=2$) with White Gaussian Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB

	Improvement in segmental SNR(dB)					PESQ score				
Noisy	-10	-5	0	5	10	1.16	1.60	2.03	2.48	2.96
MM	9.35	7.50	5.93	4.61	3.43	1.85	2.26	2.69	3.03	3.34
OM	10.59	8.65	6.65	4.80	3.26	1.73	2.16	2.56	2.91	3.23
FK	11.52	9.45	7.01	4.56	2.33	1.98	2.34	2.67	2.98	3.29
KM	11.95	10.05	7.60	4.95	2.56	1.97	2.30	2.62	2.94	3.21
SK	12.23	10.32	8.15	5.63	3.02	2.05	2.42	2.78	3.13	3.38
MSK	13.35	11.03	8.63	6.28	3.52	2.25	2.59	2.93	3.21	3.45

Table 4.8: Segmental SNR improvement and PESQ of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($(p,q)=(16,5)$, $IT=4$); KM: Full-band Kalman post-masking ($(p=10)$, $IT=4$); SK: Subband Kalman ($(p,q)=(1,1)$, $\mathbb{B}=16$, $IT=2$); and MSK: Our proposed Method ($(p,q)=(1,1)$, $\mathbb{B}=16$, $IT=2$) with Colored (Volvo) Noise, Noisy Input Segmental SNR: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB

	Improvement in segmental SNR(dB)					PESQ score				
Noisy	-10	-5	0	5	10	2.68	3.07	3.43	3.73	3.98
MM	14.08	11.86	9.72	7.50	6.03	3.06	3.53	3.79	3.98	4.10
OM	15.30	12.69	10.32	8.25	6.30	3.01	3.42	3.68	3.94	4.05
FK	16.21	13.17	10.21	7.69	5.42	3.10	3.58	3.85	4.06	4.17
KM	16.55	13.48	10.53	7.93	5.61	3.07	3.48	3.76	3.98	4.09
SK	16.86	13.92	11.15	8.53	6.38	3.21	3.65	3.89	4.07	4.16
MSK	17.12	14.52	11.70	9.02	6.56	3.29	3.72	4.00	4.16	4.25

Table 4.9: MOS score of Speech Enhanced By MM: E-M MMSE; OM: Over-subtraction masking; FK: Full-band Kalman ($(p,q)=(16,5)$, $IT=4$); KM: Full-band Kalman post-masking ($(p=10)$, $IT=4$); SK: Subband Kalman ($(p,q)=(1,1)$, $\mathbb{B}=16$, $IT=2$); and MSK: Our proposed Method ($(p,q)=(1,1)$, $\mathbb{B}=16$, $IT=2$) with Colored (Volvo) Noise, Noisy Input Segmental SNR: -5 dB with MOS = 2.18

Methods	MM	OM	FK	KM	SK	MSK
MOS	3.03	2.62	2.95	2.75	3.21	3.38

Figure 4.8: Speech spectrograms: (a) Clean speech; (b) Noisy speech with input segmental SNR = 3.5 dB (white Gaussian noise); (c) Spectral Subtraction; (d) Over-subtraction masking method; (e) E-M MMSE method; (f) Full-band Kalman filtering; (g) Subband Kalman filtering method; (h) Proposed masking-based subband Kalman method.

Figure 4.9: Speech spectrograms: (a) Clean speech; (b) Noisy speech with input segmental SNR = -10.0 dB (Volvo car interior noise); (c) Spectral Subtraction; (d) Over-subtraction masking method; (e) Full-band Kalman filtering; (f) Full-band Kalman post-masking; (g) Subband Kalman filtering method; (h) Proposed masking-based subband Kalman method.

4.6 Summary

In this chapter, an effective speech enhancement scheme has been examined by making use of the auditory masking properties, subband decomposition and Kalman filtering. A pseudo-QMF is used to implement subband decomposition and a relation between the masking threshold and the noise variance estimate used in Kalman filtering is established. Kalman filtering speech enhancement in the subband domain not only has a lower computational complexity, but also a higher SNR than that when it is implemented in the full-band domain.

As the quality of the processed speech obtained by means of Kalman filtering depends very much on the AR parameters, accurate estimation of these parameters is critical to full-band as well as subband Kalman filter design. An iterative scheme for the estimation of the AR parameters is adopted and the effect of different numbers of iterations of the proposed scheme is analyzed. The scheme is compared with the conventional full-band Kalman filtering method. We have also investigated different combinations of parameter settings for our proposed enhancement system through computer simulations. The objective of the study is to develop a high quality speech enhancement system which optimizes the trade-off between speech distortion and noise reduction. The incorporation of masking properties in the proposed scheme is found to result in significant performance improvement over the conventional subband Kalman filtering scheme. Through a large number of simulations, the advantages of the proposed system is shown in terms of objective and subjective measurements as compared to many existing speech enhancement methods.

Chapter 5

Subspace Method

5.1 Introduction

In addition to the masking-based β -order MMSE method and the subband Kalman filtering method for speech enhancement studied in the previous chapters, another category of methods that have also attracted the interests of researchers involves the use of subspace techniques. In this chapter a subspace speech enhancement method that incorporates masking properties is developed. Subspace-based approach has been shown to be largely useful for reducing wide-band additive noise. The underlying principle of the subspace method is to decompose the vector space of a noisy signal into a signal-plus-noise subspace and a noise subspace. Enhancement is then performed by removing the noise subspace and estimating the clean speech from the remaining signal-plus-noise subspace [117].

Though subspace methods have been shown to have a number of advantages in speech enhancement, psychoacoustic properties of the human auditory system have not been sufficiently exploited in these methods to date. Masking properties that researchers use for speech signal processing such as coding and enhancement are usually exploited through the critical frequency domain. In [121], a perceptual weighting filter is incorporated in a subspace approach to shape the residual noise. The perceptual weighting filter only reflects very limited properties of the human auditory system as compared to the masking threshold which is more encompassing. In

[70], a spectral domain perceptual post-filter is applied to the output of the signal subspace filter to improve the subjective quality of the enhanced speech signal. However, it cannot help to preserve the weak speech components which have already been erased by the conventional subspace method. Since the masking properties are related to the critical frequency band, which is derived from the characteristics of human cochlea, they could not be used directly in signal subspace techniques that do not function in the frequency domain. In [21], a frequency eigendomain transformation (FET) is introduced. It computes the power spectral density with respect to the estimated eigenvalues and eigenvectors according to Blackman-Tuckey cross-spectral analysis with a Bartlett window. It allows the masking properties to be incorporated into a signal subspace method. The masking-based subspace speech enhancement method is reported to yield an improved result over the conventional subspace approaches [21]. It is improved and extended to colored noise in [22]. Nevertheless, the power spectrum solution proposed in [21, 22] is a smeared version of the power spectral density (PSD) of the original signal. Subsequently, the inverse equation for eigenvalues is only an approximate solution. In this chapter, we propose an invertible solution for FET, which correctly reflects the true quantity of the eigenvalues transformed from the corresponding PSD or masking threshold [9].

Furthermore, in the conventional subspace speech enhancement methods that use perceptual properties, the audible and inaudible components are neither explicitly separated nor individually analyzed, which is in contrast to the spectral suppression approach proposed in [15]. Based on the knowledge of the human auditory system, we emphasize the suppression of the audible noise quantity in the eigendomain, and relax the extent of reduction of inaudible noise so that some weak components of the underlying speech signal will be effectively preserved and the distortion of the speech signal is therefore reduced [11]. We derive the eigen-decomposition of the autocorrelation matrix of the estimated speech with the assumption of white noise. We then propose an audible noise reduction scheme based on a signal subspace technique and outline the implementation of our proposed scheme. Finally, we further extend the scheme to the colored noise case.

5.2 Subspace Speech Enhancement

The main procedure of the subspace technique is to decompose the noisy signal into a signal-plus-noise subspace and a noise subspace by applying the Karhunen-Loève transform (KLT), modify the KLT components by a gain function and then obtain the enhanced signal by inverse KLT of the modified components [117].

A noisy speech sequence $\mathbf{x}(l)=[x(\varpi l-P+1), x(\varpi l-P+2), \dots, x(\varpi l-1), x(\varpi l)]^T$ is assumed to be the sum of a clean speech sequence $\mathbf{s}(l)=[s(\varpi l-P+1), s(\varpi l-P+2), \dots, s(\varpi l-1), s(\varpi l)]^T$ and a white noise sequence $\mathbf{w}(l)=[w(\varpi l-P+1), w(\varpi l-P+2), \dots, w(\varpi l-1), w(\varpi l)]^T$ with variance σ_w^2 , i.e.,

$$\mathbf{x}(l) = \mathbf{s}(l) + \mathbf{w}(l) \quad (5.1)$$

where l , ϖ and P denote the frame number, frame advance size and the vector dimension respectively. Let R_s denote the autocorrelation matrix of clean speech. The eigen-decomposition of R_s is given by

$$R_s = U\Lambda_s U^T \quad (5.2)$$

where $\Lambda_s = \text{diag}[\lambda_{s,0}, \dots, \lambda_{s,P-1}]$ and $U = [\mathbf{u}_0, \dots, \mathbf{u}_{P-1}]$ are the eigenvalue matrix and the orthonormal eigenvector matrix of R_s respectively. In general, there is strong empirical evidence that the autocorrelation matrices of most clean speech vectors have some eigenvalues which are practically zero. As analyzed in [117], a speech signal may be modeled as a linear combination of basis vectors, such as damped sinusoids. If the number of basis vectors is smaller than the dimension of the vector, which is usually the case, then the covariance matrix of the vector must have zero eigenvalues. The fact that some of the eigenvalues of the clean speech signal may be very close to zero indicates that the energy of the clean signal vector is distributed among a subset of its coordinates, and the signal is confined to a subspace of the noisy space. Since all eigenvalues of the white noise are strictly positive, the noise signal fills the entire vector space of the noisy signal. Hence, the vector space of a noisy signal is composed of a signal-plus-noise subspace and a complementary noise subspace. Accordingly, we need only to apply speech enhancement processing on the signal-plus-noise subspace [117], while simply discard the contents of the noise

subspace.

Assuming that the rank of the autocorrelation matrix, R_s , of a clean speech signal is M , then it has M positive eigenvalues and $P - M$ zero eigenvalues, i.e.,

$$\lambda_{s,p} = \begin{cases} \lambda_{x,p} - \sigma_w^2, & \text{if } p = 0, \dots, M-1 \\ 0, & \text{if } p = M, \dots, P-1. \end{cases} \quad (5.3)$$

Let H be a linear filter of the subspace speech enhancement method, according to the spectral domain constrained (SDC) estimator proposed by Ephraim and Van Trees in [117], we have

$$H = U \begin{bmatrix} G_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U^T \quad (5.4)$$

where G_1 is a full rank $M \times M$ matrix. Using G to substitute $\begin{bmatrix} G_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, we have $H = UGU^T$.

A possible solution of the SDC estimator is that G is a diagonal matrix given by

$$G = \Lambda_s (\Lambda_s + \sigma_w^2 \Lambda_\mu)^{-1}, \quad (5.5)$$

where $\Lambda_\mu = \text{diag}[\mu_0, \dots, \mu_{P-1}]$ is a diagonal matrix of Lagrange multipliers when deriving the estimator H based on the criterion of minimizing the speech signal distortion with a permissible residual noise level [117]. The p th diagonal element of G is given by

$$g_p = \begin{cases} \frac{\lambda_{s,p}}{\lambda_{s,p} + \mu_p \sigma_w^2}, & p = 0, \dots, M-1 \\ 0, & p = M, \dots, P-1. \end{cases} \quad (5.6)$$

5.3 Proposed Invertible Frequency Eigendomain Transformation

We can use auditory masking effects in noise suppression to overcome the classic tradeoff between noise reduction and speech distortion, wherein the audible noise is masked rather than being suppressed further, and thereby reducing chances of further distortion to speech. In order to introduce masking properties into the signal subspace field, one way is to map the masking threshold from the original frequency domain into the eigendomain.

The relation between eigenvalue and the power spectral density (PSD) can be derived through the autocorrelation function. In practice, speech signals are analyzed on a short-term basis. A short-term PSD definition through Bartlett windowed autocorrelation is given as follows:

$$\psi(k) = \sum_{p=-P+1}^{P-1} \left(1 - \frac{|p|}{P}\right) r(p) e^{-j2\pi pk/K}, \quad k = 0, \dots, K-1 \quad (5.7)$$

where K denotes the number of frequency bins, and $r(p)$ is the autocorrelation function with lag p . Based on the above definition, the following relation is satisfied [21, 72]:

$$\psi(k) = \frac{1}{P} \sum_{p=0}^{P-1} \lambda_p |V_p(k)|^2, \quad k = 0, \dots, K-1 \quad (5.8)$$

where $V_p(k)$ is the Fourier transform of unit norm eigenvector, $\mathbf{u}_p = [u_p(0) \ u_p(1), \dots, u_p(P-1)]^T$, of the autocorrelation matrix in the k th bin, which is defined by

$$V_p(k) = \sum_{i=0}^{P-1} u_p(i) e^{-j2\pi ki/K}, \quad k = 0, \dots, K-1, \quad p = 0, \dots, P-1. \quad (5.9)$$

When $K \geq 2P+1$, the inverse transformation of (5.7) is given by

$$r(p) = \frac{1}{K} \left(\frac{P}{P-p}\right) \sum_{k=0}^{K-1} \psi(k) e^{j2\pi kp/K}, \quad p = 0, \dots, P-1. \quad (5.10)$$

The inverse operation can be verified by substituting (5.7) into (5.10), i.e.,

$$\begin{aligned} r(p) &= \frac{1}{K} \left(\frac{P}{P-|p|}\right) \sum_{k=0}^{K-1} \left\{ \sum_{b=1-P}^{P-1} \left(1 - \frac{|b|}{P}\right) r(b) e^{-j2\pi kb/K} \right\} e^{j2\pi kp/K} \\ &= \frac{1}{K} \left(\frac{P}{P-|p|}\right) \sum_{b=1-P}^{P-1} \left(1 - \frac{|b|}{P}\right) r(b) \sum_{k=0}^{K-1} e^{-j2\pi(b-p)k/K} \\ &= \frac{1}{K} \left(\frac{P}{P-|p|}\right) \sum_{b=1-P}^{P-1} \left(1 - \frac{|b|}{P}\right) r(b) [K\delta(p-b)] \\ &= \frac{1}{K} \left(\frac{P}{P-|p|}\right) \left(1 - \frac{|p|}{P}\right) r(p) K = r(p), \quad 1-P \leq p \leq P-1 \end{aligned} \quad (5.11)$$

where we use the properties of periodic function:

$$\sum_{k=0}^{K-1} e^{-j2\pi(b-p)k/K} = \begin{cases} K, & b = p \\ 0, & b \neq p. \end{cases} \quad (5.12)$$

Considering the definition of eigenvalue as well as (5.10), we have

$$\begin{aligned}
\lambda_p &= \mathbf{u}_p^H R \mathbf{u}_p \\
&= \sum_{m=0}^{P-1} \sum_{b=0}^{P-1} u_p^*(m) r(m-b) u_p(b) \\
&= \sum_{m=0}^{P-1} \sum_{b=0}^{P-1} u_p^*(m) u_p(b) \frac{1}{K} \left(\frac{P}{P-|m-b|} \right) \sum_{k=0}^{K-1} \psi(k) e^{j2\pi(m-b)k/K} \\
&= \frac{1}{K} \sum_{k=0}^{K-1} \psi(k) \sum_{m=0}^{P-1} \sum_{b=0}^{P-1} u_p^*(m) u_p(b) \left(\frac{P}{P-|m-b|} \right) e^{j2\pi(m-b)k/K} \\
&= \frac{1}{K} \sum_{k=0}^{K-1} \psi(k) \left\{ \sum_{m=0}^{P-1} u_p^*(m) u_p(m) + \sum_{m=0}^{P-2} \sum_{b=m+1}^{P-1} u_p^*(m) u_p(b) \left(\frac{2P}{P+m-b} \right) \cos[2\pi(m-b)k/K] \right\}
\end{aligned} \tag{5.13}$$

where we use $u_p^*(m)u_p(b) = u_p(m)u_p^*(b) = u_p(m)u_p(b)$ as the given signal is real valued. Rewriting (5.13), the inverse transformation of eigenvalue of (5.8) can be expressed as

$$\lambda_p = \frac{1}{K} \sum_{k=0}^{K-1} \psi(k) \Omega_p(k), \quad p = 0, \dots, P-1 \tag{5.14}$$

where $\Omega_p(k)$ is given by

$$\Omega_p(k) = \sum_{m=0}^{P-1} u_p^*(m) u_p(m) + \sum_{m=0}^{P-2} \sum_{b=m+1}^{P-1} u_p^*(m) u_p(b) \left(\frac{2P}{P+m-b} \right) \cos[2\pi(m-b)k/K]. \tag{5.15}$$

Thus we give the following lemma:

- Lemma 1: The transformation pair of eigenvalue and PSD are given by (5.8) and (5.14) with $V_p(k)$ and $\Omega_p(k)$ being the functions of eigenvectors given by (5.9) and (5.15) respectively.

Fig. 5.1 (a) shows a 256-sample speech signal waveform with 8 kHz sampling rate and Fig. 5.1 (b) gives the power spectral density of the speech signal and its corresponding masking threshold. By using the conventional FET [21] and the invertible FET, Fig. 5.2 (a) gives the eigenvalues of the signal shown in Fig. 5.1. It can be seen that the eigenvalues obtained by using the invertible FET is exactly the same as the original eigenvalues, and different from the ones obtained by using the conventional FET.

In masking-based subspace speech-enhancement application, provided that the eigenvalues of speech signal have been estimated, the PSD of speech signal can be obtained through (5.8). Thus the masking threshold, $\Psi(k)$, can be computed in the frequency domain [55].

Figure 5.1: (a) Waveform of a speech signal and (b) the masking threshold.

Figure 5.2: Comparison of the original eigenvalues and the eigenvalues obtained by using our invertible FET and the conventional FET (a) the eigenvalues given the PSD and the eigenvector of the speech signal according to the FETs; (b) the masking-eigenvalues obtained by using the FETs as well as the eigenvalues obtained by the eigen-decomposition through the masking threshold.

Figure 5.3: The eigenvectors obtained by using: (o) speech signal and (*) the masking threshold.

We define the masking-eigenvalues, λ_t , as a function of masking threshold according to the frequency eigendomain transformation. Using (5.14), we have the masking-eigenvalue

$$\lambda_{t,p} = \frac{1}{K} \sum_{k=0}^{K-1} \Psi(k) \Omega_p(k), \quad p = 0, \dots, P-1. \quad (5.16)$$

In (5.16), the $\Omega_p(k)$ comes from the eigenvectors of the speech signal. As a result, the masking-eigenvalues, $\lambda_{t,p}$, can be interpreted as the transform of the masking threshold on the eigenvectors of the speech signal, i.e., they are the masking components distributed along the eigenvectors of the speech signal. As such, $\lambda_{t,p}$ can be used for the purpose of the arithmetical manipulation with the eigenvalues of speech signals in the same eigenvector space. Consequently, the masking-based subspace speech-enhancement algorithm may be realized.

The true eigenvalues, $\lambda_{mt,p}$, and eigenvectors, $\mathbf{u}_{t,p}$, of the masking threshold $\Psi(k)$ can be obtained through the eigen-decomposition of the autocorrelation matrix, R_m , of the masking threshold. The matrix, R_m , can be obtained through (5.10) as follows.

$$r_m(p) = \frac{1}{K} \left(\frac{P}{P-p} \right) \sum_{k=0}^{K-1} \Psi(k) e^{j2\pi kp/K}, \quad p = 0, \dots, P-1 \quad (5.17)$$

$$R_m = U_t \Lambda_{mt} U_t^T \quad (5.18)$$

where $R_m(i, j) = r_m(|i - j|)$, $\Lambda_{mt} = \text{diag}[\lambda_{mt,1}, \dots, \lambda_{mt,P-1}]$ and $U_t = [\mathbf{u}_{t,0}, \dots, \mathbf{u}_{t,P-1}]$. Obviously, $\lambda_{mt,p}$ are different from $\lambda_{t,p}$. In fact, $\lambda_{mt,p}$ represent the principle components of the masking threshold in its corresponding eigenvector space. In other words, they are the statistically uncorrelated components transformed from masking threshold to its corresponding eigen space. Fig. 5.2 (b) gives the masking-eigenvalues computed by using the conventional and invertible FETs as well as the corresponding eigenvalues of the masking threshold in Fig. 5.1. The eigenvectors, U_t , corresponding to the masking threshold are different from those of the speech signal. Generally, the masking-eigenvalues are not really the eigenvalues corresponding to the masking threshold. Fig. 5.3 shows some of the eigenvectors of the masking threshold and the speech signal, respectively, in Fig. 5.1.

5.3.1 Proposed Subspace-Based Audible Noise Reduction

According to psychoacoustics theory, a speech signal can mask the coexisting noise components. The noise degradation perceived by listeners will vary in time according to the time-varying properties of short-time speech components. Inspired by the spectral domain audible noise reduction technique proposed by Tsoukalas et al in [14], we derive the audible noise reduction in eigen domain. It is the audible noise components of the degradation that must be removed by the enhancement algorithm. Consequently, the enhancement approach adopted here is based on the definition of an audible noise component, which is extended and used for the derivation of an optimal modifier that achieves audible noise suppression.

Since the estimated speech signal is given by $\hat{\mathbf{s}}(l) = H\mathbf{x}(l)$, we can derive the autocorrelation

matrix of the estimated speech signal, $R_{\hat{s}}$, as follows:

$$\begin{aligned}
R_{\hat{s}} &= E\{\hat{\mathbf{s}}(l)\hat{\mathbf{s}}(l)^T\} \\
&= E\{H\mathbf{x}(l)\mathbf{x}(l)^T H^T\} \\
&= E\{H[\mathbf{s}(l) + \mathbf{w}(l)][\mathbf{s}(l) + \mathbf{w}(l)]^T H^T\} \\
&= H(R_s + \sigma_w^2 \mathbf{I})H^T \\
&= UGU^T(R_s + \sigma_w^2 \mathbf{I})UG^T U^T \\
&= UGU^T(U\Lambda_s U^T + U\sigma_w^2 U^T)UG^T U^T \\
&= UG(\Lambda_s + \sigma_w^2 \mathbf{I})G^T U^T \\
&= U\Lambda_{\hat{s}}U^T.
\end{aligned} \tag{5.19}$$

Substituting (5.5) into (5.19), we have $\Lambda_{\hat{s}}$ given by

$$\Lambda_{\hat{s}} = \Lambda_s(\Lambda_s + \sigma_w^2 \Lambda_\mu)^{-1}(\Lambda_s + \sigma_w^2 \mathbf{I})[\Lambda_s(\Lambda_s + \sigma_w^2 \Lambda_\mu)^{-1}]^T. \tag{5.20}$$

Obviously, $\Lambda_{\hat{s}}$ is a diagonal matrix with its p th diagonal element $\lambda_{\hat{s},p}$ given by

$$\lambda_{\hat{s},p} = \frac{\lambda_{s,p}^2}{(\lambda_{s,p} + \mu_p \sigma_w^2)^2}(\lambda_{s,p} + \sigma_w^2). \tag{5.21}$$

Since U is the eigenvector matrix of R_s that is the Hermitian, its vector components $(\mathbf{u}_0, \dots, \mathbf{u}_{P-1})$ are orthogonal. Hence, by the definition of eigen-decomposition, we have the following lemma:

- Lemma 2: The eigenvalue of the estimated speech is $\lambda_{\hat{s},p}$ given by (5.21)
- Lemma 3: The eigenvector of the estimated speech is the same as that of clean speech.

Lemmas 2 and 3 can be proven as follows:

Since the autocorrelation matrix, R_s , is Hermitian, i.e., $R_s^H = R_s$, its eigenvectors are orthogonal to each other, i.e., $UU^T = \mathbf{I}$. As a result, we have

$$\begin{aligned}
R_{\hat{s}} - \lambda_{\hat{s},p} \mathbf{I} &= U\Lambda_{\hat{s}}U^T - \lambda_{\hat{s},p}UU^T \\
&= U(\Lambda_{\hat{s}} - \lambda_{\hat{s},p} \mathbf{I})U^T, \quad p = 0, \dots, P-1.
\end{aligned} \tag{5.22}$$

Using the distributive property of determinant, the determinant of the above matrix can be written as

$$\begin{aligned}
 |R_{\hat{s}} - \lambda_{\hat{s},p}\mathbf{I}| &= |U| |\Lambda_{\hat{s}} - \lambda_{\hat{s},p}\mathbf{I}| |U^T| = |U| |\Lambda_{\hat{s}} - \lambda_{\hat{s},p}\mathbf{I}| \frac{1}{|U|} \\
 &= |\Lambda_{\hat{s}} - \lambda_{\hat{s},p}\mathbf{I}| \\
 &= \begin{vmatrix} \ddots & \vdots & \vdots & \vdots & \ddots \\ \cdots & \lambda_{\hat{s},p-1} - \lambda_{\hat{s},p} & 0 & 0 & \cdots \\ \mathbf{0} & 0 & 0 & 0 & \mathbf{0} \\ \cdots & 0 & 0 & \lambda_{\hat{s},p+1} - \lambda_{\hat{s},p} & \cdots \\ \ddots & \vdots & \vdots & \vdots & \ddots \end{vmatrix} \\
 &= 0.
 \end{aligned} \tag{5.23}$$

(5.23) implies that $\lambda_{\hat{s},p}$ is the eigenvalue of $R_{\hat{s}}$. According to (5.19), $\mathbf{u}_p$ is the eigenvector of $R_{\hat{s}}$ corresponding to $\lambda_{\hat{s},p}$, i.e.,

$$R_{\hat{s}} \mathbf{u}_p = \lambda_{\hat{s},p} \mathbf{u}_p, \quad p = 0, \dots, P-1. \tag{5.24}$$

Thus, we prove that $\lambda_{\hat{s},p}$ and $\mathbf{u}_p$ are the eigenvalue and corresponding eigenvector of the estimated speech signal.

According to (5.2), and with the fact that $R_x = R_s + \sigma_w^2 \mathbf{I}$, we have

$$\begin{aligned}
 R_x &= U \Lambda_x U^T, \\
 \Lambda_x &= \Lambda_s + \sigma_w^2 \mathbf{I}
 \end{aligned} \tag{5.25}$$

where $\Lambda_x = \text{diag}[\lambda_{x,0}, \dots, \lambda_{x,P-1}]$ is the eigenvalue matrix of the noisy signal. Subsequently, Lemma 3 is extended to Lemma 3a as follows:

- Lemma 3a: The eigenvector matrix of the estimated speech obtained by using the SDC subspace estimator is the same as those of the clean speech and the noisy speech.

This means that the corresponding eigenvalues of the estimated speech, clean speech and the noisy speech can be compared and manipulated arithmetically in the same eigenvector space.

The main objective of psychoacoustics based speech enhancement is to reduce the audible noise components. Provided that the masking threshold of the clean speech signal is known, those

noisy components below this threshold is inaudible due to the effect of masking. The modification of the noisy signal may be based on the rule that it should affect only the observed speech signal to the degree of the audible level of the noise such that the distortion introduced to the original speech signal is reduced. Next, we define the audible eigenvalues of the autocorrelation matrices of the clean speech and the estimated speech as the maximum values of their individual eigenvalues and the corresponding masking-eigenvalue respectively, i.e.,

$$\vartheta_{s,p} = \max(\lambda_{s,p}, \lambda_{t,p}) \quad (5.26)$$

$$\vartheta_{\hat{s},p} = \max(\lambda_{\hat{s},p}, \lambda_{t,p}). \quad (5.27)$$

Subsequently, we define the error between the two audible eigenvalues in (5.26) and (5.27) as the residual audible eigen-noise, i.e.,

$$\varepsilon(p) = \vartheta_{\hat{s},p} - \vartheta_{s,p}, \quad p = 0, \dots, P-1. \quad (5.28)$$

Substituting (5.26) and (5.27) into (5.28), the residual audible eigen-noise can be expressed as

$$\varepsilon(p) = \begin{cases} \lambda_{\hat{s},p} - \lambda_{s,p}, & \text{if } \lambda_{\hat{s},p} > \lambda_{t,p} \text{ \& } \lambda_{s,p} > \lambda_{t,p} \text{ (I)} \\ \lambda_{\hat{s},p} - \lambda_{t,p}, & \text{if } \lambda_{\hat{s},p} > \lambda_{t,p} \text{ \& } \lambda_{s,p} \leq \lambda_{t,p} \text{ (II)} \\ \lambda_{t,p} - \lambda_{s,p}, & \text{if } \lambda_{\hat{s},p} \leq \lambda_{t,p} \text{ \& } \lambda_{s,p} > \lambda_{t,p} \text{ (III)} \\ 0, & \text{if } \lambda_{\hat{s},p} \leq \lambda_{t,p} \text{ \& } \lambda_{s,p} \leq \lambda_{t,p} \text{ (IV)}. \end{cases} \quad (5.29)$$

Since we only have the noisy speech available, and the estimated speech is obtained through modifying the noisy signal, the conditions of (5.29) can be extended to the comparisons between the eigenvalues of the autocorrelation matrices of the noisy and clean speech signals and the masking-eigenvalue. Therefore, we further define the modified residual audible eigen-noise as

$$\tilde{\varepsilon}(p) = \begin{cases} \lambda_{\hat{s},p} - \lambda_{s,p}, & \text{if } \lambda_{x,p} > \lambda_{t,p} \text{ \& } \lambda_{s,p} > \lambda_{t,p} \text{ (I)} \\ \lambda_{\hat{s},p} - \lambda_{t,p}, & \text{if } \lambda_{x,p} > \lambda_{t,p} \text{ \& } \lambda_{s,p} \leq \lambda_{t,p} \text{ (II)} \\ \lambda_{t,p} - \lambda_{s,p}, & \text{if } \lambda_{x,p} \leq \lambda_{t,p} \text{ \& } \lambda_{s,p} > \lambda_{t,p} \text{ (III)} \\ 0, & \text{if } \lambda_{x,p} \leq \lambda_{t,p} \text{ \& } \lambda_{s,p} \leq \lambda_{t,p} \text{ (IV)}. \end{cases} \quad (5.30)$$

In fact, the main difference between $\tilde{\varepsilon}(p)$ and the discrepancy between the eigenvalues of noisy and clean speech signals is that the former has a smaller dynamic range than the latter as

a result of involving $\lambda_{t,p}$. Consequently, it results in more emphasis of audible noise in the processing algorithm. As an eigenvalue represents the power component along its corresponding eigenvector and, in the additive white noise case, the eigenvalue for noisy signal is the sum of the corresponding eigenvalues of clean speech and noise, it is therefore always greater than that of clean speech. The goal of our proposed enhancement algorithm is to modify the noisy eigenvalue by attenuating it to obtain the estimated eigenvalue for clean speech. Thus, the objective of our proposed method is to make the modified residual audible eigen-noise less than or equal to zero, i.e.,

$$\tilde{\varepsilon}(p) \leq 0, \quad p = 0, \dots, P-1. \quad (5.31)$$

When $\tilde{\varepsilon}(p)$ is negative, it means that either the speech eigenvalue is underestimated (condition of (5.30, I)) or the speech eigenvalue is correctly estimated to be below the threshold (condition of (5.30, II)) and is hence inaudible. Observing (5.30), we may give the following conclusions for each condition:

- (5.30, I and II): $\tilde{\varepsilon}(p)$ may be positive, negative or zero, depending on the values of $\lambda_{\hat{s},p}$, $\lambda_{s,p}$ and $\lambda_{t,p}$
- (5.30, III): $\tilde{\varepsilon}(p)$ is always negative
- (5.30, IV): $\tilde{\varepsilon}(p)$ is zero.

Consequently, (5.30, III) and (5.30, IV) always satisfy the criterion in (5.31), wherein $\lambda_{x,p} \leq \lambda_{t,p}$. For the cases of (5.30, I) and (5.30, II), an audible noise reduction constraint needs to be applied in order to satisfy the criterion in (5.31), wherein $\lambda_{x,p} > \lambda_{t,p}$. Considering (5.6), it is noted that the Lagrange multipliers play an important role on the estimation of speech signal. The value of μ_p regulates the trade-off between the residual noise and speech distortion. As discussed in [117], when μ_p increases from zero to infinity, the level of the residual noise decreases whilst the level of signal distortion increases. An approach for choosing μ_p that compromises between signal distortion and residual noise is to make μ dependent on the SNR of each frame [114]. A simpler approach is to set μ_p to a fixed value in the time domain constrain (TDC) method proposed in

[117], but it leads to a suboptimal selection of μ_p . To overcome this drawbacks, we propose to adapt the value of μ_p based on the criterion of audible noise reduction defined in (5.31).

As indicated by the above analysis, the condition of (5.31) is always satisfied in the case of $\lambda_{x,p} \leq \lambda_{t,p}$ and we should therefore keep the eigenvalue unchanged, i.e., $\lambda_{\hat{s},p} = \lambda_{x,p}$. According to (5.21), we then have

$$\frac{\lambda_{s,p}^2}{(\lambda_{s,p} + \mu_p \sigma_w^2)^2} (\lambda_{s,p} + \sigma_w^2) = \lambda_{x,p}. \quad (5.32)$$

Since $\lambda_{x,p} = \lambda_{s,p} + \sigma_w^2$ for white additive noise, it can be seen that $\mu_p = 0$.

In the case of $\lambda_{x,p} > \lambda_{t,p}$, and according to the constraint of (5.31), substituting (5.21) into (5.30, I) and (5.30, II) respectively yields

$$\begin{aligned} \mu_{I,p} &\geq \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{s,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2}, & \lambda_{s,p} > \lambda_{t,p} \\ \mu_{II,p} &\geq \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2}, & \lambda_{s,p} \leq \lambda_{t,p} \end{aligned} \quad (5.33)$$

where $\mu_{I,p}$ and $\mu_{II,p}$ represent the μ_p value in (5.30, I) and (5.30, II), respectively. In order to find a solution of μ_p to satisfy the two inequations of (5.33), we use the masking-eigenvalue in place of clean speech eigenvalue ($\lambda_{s,p}$) required in the square root part of (5.33), i.e., we use $\mu_{t,p} = \left(\sqrt{\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2}$ to replace both $\mu_{I,p}$ and $\mu_{II,p}$. It is found that when $\lambda_{s,p} > \lambda_{t,p}$, $\left(\sqrt{\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) > \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{s,p}}} - 1 \right) \geq 0$; and when $\lambda_{s,p} \leq \lambda_{t,p}$, $\left(\sqrt{\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) \geq \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right)$. As R_s is a nonnegative matrix, $\lambda_{s,p}$ is a nonnegative value, i.e., $\lambda_{s,p} \geq 0$, and $\sigma_w^2 \geq 0$, we therefore have

$$\mu_{t,p} \geq \begin{cases} \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{s,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2}, & \lambda_{s,p} > \lambda_{t,p} \\ \left(\sqrt{\frac{\lambda_{s,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2}, & \lambda_{s,p} \leq \lambda_{t,p}. \end{cases} \quad (5.34)$$

As a result, $\mu_{t,p}$ satisfies both the conditions of (5.33) simultaneously. In other words, it implies that $\mu_{t,p}$ satisfies the criterion of (5.31) under both the conditions of (5.30, I) and (5.30, II) simultaneously. Thus the estimate of μ_p may be determined by

$$\hat{\mu}_p = \left(\sqrt{\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}}} - 1 \right) \frac{\lambda_{s,p}}{\sigma_w^2} + \delta \quad (5.35)$$

where δ is a nonnegative value, which is empirically set to $2.1 \sim 3.8$, and the noise variance, σ_w^2 , is estimated during a speech-absence period. Simulations show that using $\mu_p = \delta$ to replace

$\mu_p = 0$ when $\lambda_{x,p} < \lambda_{t,p}$ does lead to a better enhancement performance. As a conclusion, we have

$$\hat{\mu}_p = \begin{cases} (\sqrt{\frac{\lambda_{t,p} + \sigma_w^2}{\lambda_{t,p}}} - 1) \frac{\lambda_{s,p}}{\sigma_w^2} + \delta, & \lambda_{s,p} + \sigma_w^2 > \lambda_{t,p} \\ \delta, & \lambda_{s,p} + \sigma_w^2 < \lambda_{t,p}. \end{cases} \quad (5.36)$$

5.3.2 Proposed Approach for Colored Noise

For colored noise with its autocorrelation matrix R_n , we may pre-whiten the noise by multiplying $R_n^{-\frac{1}{2}}$ with the observed noisy signal $\mathbf{x}(l)$ so that the input noisy signal becomes $R_n^{-\frac{1}{2}}\mathbf{x}(l)$ in which the noise signal is white [31]. We then input the whitened noisy signal to the speech enhancement system. After processing, we obtain the final estimated speech by multiplying the estimated speech signal $\hat{\mathbf{s}}(l)$ by $R_n^{\frac{1}{2}}$.

The autocorrelation function of the pre-whitened noisy signal is given by

$$\begin{aligned} R_{prew_x} &= E\{R_n^{-\frac{1}{2}}\mathbf{x}(l)\mathbf{x}^T(l)(R_n^{-\frac{1}{2}})^T\} \\ &= R_n^{-\frac{1}{2}}E\{\mathbf{x}(l)\mathbf{x}^T(l)\}(R_n^{-\frac{1}{2}})^T \\ &= R_n^{-\frac{1}{2}}R_x(R_n^{-\frac{1}{2}})^T \\ &= R_n^{-\frac{1}{2}}(R_s + R_n)(R_n^{-\frac{1}{2}})^T \\ &= R_n^{-\frac{1}{2}}R_s(R_n^{-\frac{1}{2}})^T + \mathbf{I} \\ &= R_{prew_s} + \mathbf{I}. \end{aligned} \quad (5.37)$$

and the estimated speech is given by the following equation:

$$\hat{\mathbf{s}} = R_n^{\frac{1}{2}}UG_{prew}U^TR_n^{-\frac{1}{2}}\mathbf{x}(l). \quad (5.38)$$

where $G_{prew} = \Lambda_{prew_s}(\Lambda_{prew_s} + \Lambda_\mu)^{-1}$, and Λ_{prew_s} is the eigenvalue matrix of R_{prew_s} . Here, the pre-whitening method has changed the focus from the original speech signal into a pre-whitened clean speech signal. In other words, the original optimization constraint used for minimizing the variance of speech distortion is shifted to minimizing the variance of pre-whitened speech distortion. Therefore, it may not result in the optimal estimation of the clean speech signal.

In [120], a generalized subspace approach is proposed for enhancing speech corrupted by colored noise. The orthogonal KLT used for white noise is replaced by a nonsingular nonorthogonal

transformation, and the eigenvalues of the clean speech are essentially replaced by those of the whitened signal. It results in no explicit solution of the estimation of clean speech, but the whitened speech instead.

The total distortion of the estimation is given as follows

$$d = \hat{\mathbf{s}} - \mathbf{s} = H\mathbf{x} - \mathbf{s} = (H - I)\mathbf{s} + H\mathbf{w} = d_s + d_n \quad (5.39)$$

where $d_s = (H - I)\mathbf{s}$ is the signal distortion and $d_n = H\mathbf{w}$ the residual noise. Let $\epsilon_s = \text{tr}\{\mathbf{E}\{d_s d_s^T\}\}$. In order to provide proper shaping of the spectrum of the residual noise and minimization of the variance of the signal distortion, a constrained optimization problem is proposed in [107], i.e.,

$$\min_H \epsilon_s \quad (5.40)$$

subject to the constraints

$$\mathbf{E}\left\{\left|\mathbf{u}_p^T d_n\right|^2\right\} \leq \alpha_p \sigma_n^2(p), \quad p = 0, \dots, P-1 \quad (5.41)$$

where $\sigma_n^2(p)$ is the p th diagonal element of $U^T R_n U$, and $0 \leq \alpha_p \leq 1$ is the constraint constant. According to the time domain constraint estimator defined in [117],

$$\begin{aligned} H_{opt} &\triangleq \arg \left\{ \min_H \epsilon_s \right\} \\ \text{Subject to : } &\frac{1}{P} \text{tr}\{\mathbf{E}\{d_n d_n^T\}\} \leq \sigma^2 \end{aligned} \quad (5.42)$$

where σ^2 is a positive constant, and [1] gives the solution as:

$$\begin{aligned} H_{opt} &= R_s(R_s + \mu R_n)^{-1} \\ &= U\Lambda_s(\Lambda_s + \mu U^T R_n U)^{-1}U^T. \end{aligned} \quad (5.43)$$

In [1], $U^T R_n U$ is approximated by a diagonal matrix with its p th diagonal element equal to the variance of noise along the p th eigenvector of the clean speech autocorrelation matrix, i.e.,

$$U^T R_n U \simeq \begin{bmatrix} \sigma_n^2(0) & 0 & \cdots & 0 \\ 0 & \sigma_n^2(1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2(P-1) \end{bmatrix} \quad (5.44)$$

Based on the analysis of the SDC and TDC solutions proposed by [107] and [1] respectively, we propose to modify (5.5) as follows:

$$\tilde{G} = \Lambda_s(\Lambda_s + \Sigma_n \tilde{\Lambda}_\mu)^{-1} \quad (5.45)$$

$$\Sigma_n = \begin{bmatrix} \sigma_n^2(0) & 0 & \cdots & 0 \\ 0 & \sigma_n^2(1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma_n^2(P-1) \end{bmatrix} \quad (5.46)$$

where $\tilde{\Lambda}_\mu$ is a diagonal matrix with its p th element $\tilde{\mu}_p$ modified according to (5.36) as

$$\tilde{\mu}_p = \begin{cases} (\sqrt{\frac{\lambda_{t,p} + \sigma_n^2(p)}{\lambda_{t,p}}} - 1) \frac{\lambda_{s,p}}{\sigma_n^2(p)} + \tilde{\delta}, & \lambda_{s,p} + \sigma_n^2(p) > \lambda_{t,p} \\ \tilde{\delta}, & \lambda_{s,p} + \sigma_n^2(p) < \lambda_{t,p} \end{cases} \quad (5.47)$$

and $\tilde{\delta}$ is set to $1.8 \sim 3.8$. Consequently the estimate of the speech signal degraded by colored noise is obtained by

$$\hat{\mathbf{s}}(l) = U \tilde{G} U^T \mathbf{x}(l). \quad (5.48)$$

5.3.3 Implementation

White Noise

In practice, the autocorrelation function can be approximated as follows:

$$r_x(p) \approx \frac{1}{N} \sum_{i=1}^N x(i)x(i+p), \quad p = 0, \dots, P-1. \quad (5.49)$$

and the autocorrelation matrix, R_x , is defined as ¹:

$$R_x = \begin{bmatrix} r_x(0) & r_x(1) & \cdots & r_x(P-1) \\ r_x(1) & r_x(0) & \cdots & r_x(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(P-1) & r_x(P-2) & \cdots & r_x(0) \end{bmatrix} \quad (5.50)$$

With the PSD, ψ_s , of a speech signal, we can obtain the masking threshold Ψ according to [[55], pp. 317-319]. Using the method proposed in [21], the masking-eigenvalues can be

¹For real value, the complex conjugate mark is ignored.

approximately obtained based on the masking-threshold according to (5.14) by replacing λ_p and $\psi(k)$ with $\lambda_{t,p}$ and $\Psi(k)$ respectively.

Next, we give the procedure for the implementation of the proposed algorithm

1. Estimate σ_w^2 , followed by pre-estimation of R_s by $\hat{R}_s = R_x - \hat{\sigma}_w^2 \mathbf{I}$.
2. Perform eigen-decomposition of $\hat{R}_s$ according to (5.2).
3. Assuming that the eigenvalues of R_s are ordered as $\lambda_{s,0} \geq \lambda_{s,1} \geq \dots \geq \lambda_{s,P-1}$, determine the dimensionality of the speech signal subspace as follows:

$$M = \arg \max_{p \in [0, \dots, P-1]} \{\lambda_{s,p} > 0\}. \quad (5.51)$$

4. Compute $V_p(k)$ according to (5.9) and the pre-estimated PSD of speech signal $\hat{\psi}_s$ according to (5.8) with P changed into M , i.e.,

$$\hat{\psi}_s(k) = \frac{1}{P} \sum_{p=1}^M \hat{\lambda}_{s,p} |V_p(k)|^2. \quad (5.52)$$

5. Use $\hat{\psi}_s$ to obtain the masking threshold $\hat{\Psi}$.
6. Use $\hat{\Psi}$ to obtain $\lambda_{t,p}$ according to FET (the invertible FET (5.16) or the conventional one [21]).
7. Compute μ_p according to (5.35) when $\lambda_{x,p} > \lambda_{t,p}$, and $\mu_p = 0$ when $\lambda_{x,p} \leq \lambda_{t,p}$.
8. Compute g_p according to (5.6).
9. Compute H according to (5.4) and then estimate the enhanced speech signal by $\hat{\mathbf{s}}(l) = H\mathbf{x}(l)$.

Colored Noise

Next, we give the procedure for the implementation of the proposed algorithm for the case of colored noise.

1. Estimate the autocorrelation of noise, $\hat{R}_n$, followed by pre-estimation of R_s , i.e.,

$$\hat{R}_s = R_x - \hat{R}_n. \quad (5.53)$$

2. Follow step 2 through to step 6 of the white noise case.
3. For (5.46), $\sigma_n^2(p)$ is approximated to be the p th diagonal element of $U^T R_n U$, where U is the eigenvector matrix of $\hat{R}_s$.
4. Compute μ_p according to (5.47).
5. Compute $\tilde{G}$ according to (5.45) and then estimate the enhanced speech signal according to (5.48).

For comparison purpose, we also implemented the pre-whitening method for colored noise. The procedure is given as follows:

1. Estimate the autocorrelation of noise, $\hat{R}_n$.
2. Pre-whiten the noisy speech by multiplying $\hat{R}_n^{-\frac{1}{2}}$.
3. Follow step 2 through to step 9 of the white noise case.
4. Obtain the estimated speech by multiplying $\hat{R}_n^{\frac{1}{2}}$.

5.4 Performance Evaluation

For the purpose of evaluation, three different noise types taken from the NOISEX-92 database [2], they are white noise, F-16 cockpit noise and Volvo car interior noise. Twenty different speech utterances from the TIMIT database [45] are used in our performance evaluations, with ten spoken by male and another ten by female speakers. The noises and utterances were down-sampled to 8 kHz, respectively, before simulation.

In order to maintain the whiteness of the input noise, i.e., $R_n = \sigma_w^2 \mathbf{I}$, a rectangular analysis window is used [117]. The proposed enhancement parameters are set to: sampling rate = 8 kHz,

number of samples used to calculate autocorrelation = 256, $\varpi = 16$, $P = 32$ and $K = 256$. The enhanced signal is Hanning windowed and combined using the overlap-add synthesis method. The estimation of noise may be obtained during speech pause intervals. In our simulations, since we focus on speech enhancement algorithm study and only the stationary noise signal is considered, the statistical characteristics of the noise signal is directly obtained from an initial segment of the noisy signal, where only pure noise is present.

The objective measurements used for performance evaluation are: segmental SNR improvement, perceptual evaluation of speech quality (ITU-P.862) (PESQ) and modified Bark spectral distortion (MBSD). Subjective assessments include spectrogram inspection as well as informal listening tests. For comparison purpose, conventional subspace methods and the proposed subspace methods are implemented. As it is well known that segmental SNR is more accurate in indicating the speech distortion than the global SNR, noise was added at different input segmental SNRs for each of utterances in our simulations. The segmental SNR is measured by computing the SNR for each of the frames (including speech activity phase and speech pause phase) and averaging these SNRs over the entire utterance.

For the case of white noise, the methods assessed include the conventional subspace method proposed by Ephraim and Van Trees (EV) [117], a masking-based subspace method proposed by Jabloun and Champagne (JC-1) [21] and its improved version (JC-2) [22], as well as our proposed subspace-based audible noise reduction method with (i) the conventional FET and white noise model (CFE-ANR-1), and (ii) the invertible FET and white noise model (IFE-ANR-1). For the case of colored noise, the methods include the pre-whitening subspace method proposed by Lev-Ari and Ephraim (PWH-LE) [31], JC-2, our proposed subspace-based audible noise reduction method with (i) pre-whitening and invertible FET (PWH-ANR), (ii) the conventional FET and colored noise model (CFE-ANR-2) and (iii) the invertible FET and colored noise model (IFE-ANR-2).

Fig. 5.4 shows the comparison results for the case of white noise. In order to investigate the effectiveness of denoising for different colored noises, Figs. 5.5 and 5.6 show the comparison

Figure 5.4: Objective comparison for the case of white noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (—) Noisy speech (White noise), ($\square$) EV, ($\diamond$) JC-1, ($\circ$) JC-2, ($*$) CFE-ANR-1, ($+$) IFE-ANR-1.

Figure 5.5: Objective comparison for the case of F-16 (colored) noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (—) Noisy (F-16 cockpit noise), ($\square$) PWH-LE, ($\circ$) JC-2, (Δ) PWH-ANR, ($*$) CFE-ANR-2, ($+$) IFE-ANR-2.

Figure 5.6: Objective comparison for the case of Volvo (colored) noise in terms of (a) Segmental SNR; (b) PESQ score; (c) MBSD. (-) Noisy (F-16 cockpit noise), ($\square$) PWH-LE, ($\circ$) JC-2, (Δ) PWH-ANR, (*) CFE-ANR-2, (+) IFE-ANR-2.

Figure 5.7: Speech spectrogram of a clean speech.

Table 5.1: Preference results in the case of white noise

White Noise	IFE-ANR-1 to Noisy (unprocessed)	IFE-ANR-1 to EV	IFE-ANR-1 to JC-1
Male	57.5%	80.0%	72.5%
Female	70.0%	77.5%	62.5%
Average	63.75%	78.75%	67.50%
Remarks	Noisy: high intelligibility; high noise	EV: More artificial noise	JC-1: More distortion

Figure 5.8: Speech spectrograms of noisy and enhanced speech: (a) Noisy speech (White noise), (b) EV, (c) JC-1, (d) IFE-ANR-1.

Figure 5.9: Speech spectrograms of noisy and enhanced speech: (a) Noisy (F-16 cockpit noise), (b) JC-2, (c) PWH-ANR, (d) IFE-ANR-2.

results for F-16 cockpit noise and Volvo interior car noise, respectively. It can be observed that the performances of CFE-ANR-(1/2) and IFE-ANR-(1/2) are generally better than the others as expected. It is also noted that IFE-ANR-(1/2) is always better than CFE-ANR-(1/2). It shows, at least based on simulations with the range of input SNRs under consideration, that the invertible FET leads to better results than the conventional FET. Another observation is that for the case of the colored noise shown in Figs. 5.5 and 5.6, there is an apparent drop in performance arising from the use of pre-whitening. This is largely because the original characteristics of a speech signal are changed and the characteristics of the changed signal can be very much different from those of a speech signal. When using eigen-decomposition, pre-whitening may greatly weaken the advantage arising from the characteristics of principal component assumption of speech signal. Since the deviations of the eigenvectors of the pre-whitened speech from the eigenvectors of the original speech are usually large, when we discard the noise components, we might also discard more speech components due to the use of the criterion of separation of the signal and noise subspaces as well as eigen-filtering. Statistically, the pre-whitened speech is more like noise and therefore the difference between the pre-whitened speech and noise is less apparent than that between the original speech and noise. We therefore may not be able to separate the speech signal and noise signal effectively through the use of pre-whitening. This is why JC-2, CFE-ANR-2 and IFE-ANR-2 have better performances than PWH-LE and PWH-ANR.

Fig. 5.7 shows the spectrogram of a clean speech. The marked time-frequency parts A, B, C, D, E and F in the figure will be referred to for comparison in the following. Fig. 5.8 shows the spectrograms of the speech (of Fig. 5.7) contaminated by white noise and the enhanced speech signals. The subfigures (b), (c) and (d) of Fig. 5.8 are the spectrograms of the estimated speech processed by EV, JC-1 and IFE-ANR-1 respectively. The spectral components in parts A and F of the subfigures (b) and (d) are clearer than those of (c). The spectral components in part B of (d) are clearer than those in (c), and those of (b) almost disappears. Obviously, the spectral components in parts C and E of (d) are better than the corresponding areas in (b) and (c). The spectral components in part D in the three subfigures are very similar.

Fig. 5.9 (a) shows the spectrograms of the noisy speech, which is the same speech given in Fig. 5.7 but contaminated by F-16 noise. Figs. 5.9 (b), (c) and (d) are the spectrograms of the enhanced speech signals obtained by JC-2, PWH-ANR and IFE-ANR-2, respectively. From the spectrograms, it can be seen that the spectral components in part A of (d) are slightly better preserved than those of (b) and (c). The spectral components in part B of (b) and (d) are quite similar, and those of (c) disappear. The spectral components in part C of the three subfigures are very similar. In part D, the spectral components of (d) and (b) are similar but very much better than those of (c). The spectral components in part E of (d) are slightly better than those of (b), but are apparently clearer than those of (c). The components in part F of (d) are better than the corresponding ones in (c), and those of (b) almost disappear.

The aim of speech enhancement is to improve the quality of a degraded speech signal while keeping the intelligibility at a certain level. Speech quality is a subjective measure that reflects the perception of listeners. It may represent the degree of pleasantness of a speech signal and the degree of the effort needed for understanding the message of a speech signal. Intelligibility is generally used as an objective measure of the amount of information that can be extracted by listeners from a speech signal, regardless of whether the signal is clean or noisy. A group of 20 listeners participated in our listening tests; only 8 of them are researchers working in the area of acoustic signal processing. The listeners are asked to mark their preferred utterance for each given pair of speech utterances in terms of intelligibility, speech distortion and residual noise. High quality earphones of model K-240DF are provided for the listening tests. Two speech sentences contaminated by different types of noise, which include white noise, F-16 cockpit noise and Volvo interior car noise, at input segmental SNR of -5 dB, are processed by different enhancement methods. One of the two sentences, "Combine all the ingredients in a large bowl.", is spoken by a male, and the other, "In wage negotiations, the industry bargains as a unit with a single union.", by a female. Listeners are required to listen three times to each of the speech signals including noisy and enhanced ones. For each pair of utterances, the listening order of the speech signals enhanced by various enhancement methods is hidden from the listeners who only give a selection

Table 5.2: Preference results in the case of F-16 noise

F-16 Noise	IFE-ANR-2 to Noisy (unprocessed)	IFE-ANR-2 to PWH-LE	IFE-ANR-2 to JC-2
Male	77.5%	97.5%	87.5%
Female	62.5%	95.0%	82.5%
Average	70.00%	96.25%	85.00%
Remarks	Noisy: high annoying noise; high intelligibility	PWH-LE: Much distortion	Both IFT-ANR-2 and JC-2: Can feel the distortion

for each pair of utterances. For those pairs which cannot be given the more preference, we assign the same percentage of preference to both sides of each of the pairs when calculate the preference results. Tables 5.1, 5.2 and 5.3 show the preferences of the invited listeners corresponding to the cases of white noise, F-16 cockpit noise and Volvo car interior noise respectively. The informal listening test results show that most listeners prefer the enhanced speech obtained by using IFE-ANR-(1,2) as compared to EV, JC-1, PWH-LE and JC-2.

5.5 Summary

In this chapter, a subspace-based audible noise reduction method for enhancing the noise corrupted speech has been proposed. Firstly, an invertible frequency eigendomain transformation is derived. By using the transformation, the masking properties of the human auditory system can be made use of in the eigendomain. The masking properties can therefore be incorporated into a signal subspace technique through the definition of residual audible eigen-noise, wherein the gain of the eigen-filter is varied according to an audible noise reduction criterion for the white noise model. Our proposed method is further extended to the colored noise model. Moreover, we describe the implementation procedures of the proposed method. From the simulation results ob-

Table 5.3: Preference results in the case of Volvo noise

Volvo noise	IFE-ANR-2 to Noisy (unprocessed)	IFE-ANR-2 to PWH-LE	IFE-ANR-2 to JC-2
Male	52.5%	87.5%	72.5%
Female	60.0%	92.5%	77.5%
Average	56.25%	90.00%	75.00%
Remarks	Noisy: high low-frequency noise; more natural speech;	PWH-LE: Much distortion	Both IFT-ANR-2 and JC-2: Can feel the distortion

tained, it is shown that the proposed audible noise reduction method based on the invertible FET is better than that based on the conventional FET. The performance of the proposed method outperforms the conventional EV and JC's subspace methods in terms of various measurements and informal listening tests.

Chapter 6

Performance Comparison among the Proposed Speech Enhancement Methods

6.1 Introduction

In the previous three chapters, we have studied three speech enhancement approaches, namely spectral suppression, Kalman filtering and signal subspace approach, and correspondingly, proposed three masking-based enhancement methods. They are the proposed masking-based β -order MMSE (M β OM) method which is based on spectral suppression; the masking-based sub-band Kalman filtering (MSKF) method; and the subspace-based audible noise reduction (SANR) method, with both white and colored models¹. In this chapter, a detailed comparison among the three methods will be provided in order to highlight their relative strengths and weaknesses, so that suitable methods for particular scenarios may be chosen depending on specific applications.

¹Refer to the last chapter, in this chapter, we choose IFE-ANR-1 and IFE-ANR-2 for both white and colored models.

6.2 Comparison among the Enhancement Methods

In this section, we first compare the proposed M β OM, MSKF and SANR methods in terms of commonly used objective and subjective performance measurements. It is well known that segmental SNR, IS distortion measure, PESQ as well as MBSD measure have their own merits in verifying the effectiveness of an enhancement algorithm. Each of them has its own way of reflecting a certain aspect of performance of an enhancement algorithm. Segmental SNR measures the closeness of the waveform of an enhanced speech to that of the clean speech in the time domain. The IS distortion reflects the distance of the spectral envelopes of the enhanced speech and the clean speech. The PESQ is designed to assess the perceptual quality of speech signal for telephony communication. The MBSD marks the loudness distance in the Bark domain between the enhanced speech and clean speech with consideration of the masking properties of the human auditory system. In order not to over-emphasize the strength of each measure, it is necessary to be reminded of its weakness so that we may have a clearer understanding of the performance assessment of each of the speech enhancement methods. For example, the IS distortion measure reflects some linear relationship between the IS score and the degree of contamination when it is used to measure the distortion of a speech signal corrupted by additive noise. However, when it is used to evaluate the quality of an enhanced speech signal, it may not be a very fair measure. The main reason is because some singular values may be produced with the enhanced speech. In general, a speech enhancement system involves a speech enhancement procedure and noise estimation. Usually, we only emphasize the former, i.e., the speech enhancement method, which is the core of a speech enhancement system. As the IS measure depends greatly on the estimation of the noise statistical quantity, the enhancement system with different noise estimators may give very different IS values when the same speech enhancement method is used. Another example is PESQ which indicates the perceptual quality of a speech signal. Actually, PESQ does not attempt to define what quality is. The aim of the PESQ algorithm is to objectively predict the subjective “mean opinion score” in an ITU-P.800 listening setup. We believe that PESQ does what it is intended to do, but it does not include the effects of loudness loss (ITU-T P.862 Table

2). It also does not take into account the frequency response variations of less than 20 dB (ITU-T P.862 10.2.6). As a result, PESQ can be used in addition to other measurement methods when evaluating the performance of a speech enhancement system; however, a good PESQ score alone cannot be taken as the final endorsement of a speech enhancement system.

6.2.1 Characteristics of the Enhancement Methods

A. COMPUTATIONAL COMPLEXITY

Each of the three methods involves the computation of the auditory masking threshold. The computational complexity will be discussed as follows with 8000 Hz sampling rate. We use FLOPS (floating-point operations per second) to measure the computational complexity. As $M\beta OM$ uses look-up table search for hyper-functions, it requires about 20 MFLOPS (million FLOPS). With three iterations, MSKF requires about 900 and 950 MFLOPS for white and colored models respectively. SANR requires about 50 and 70 MFLOPS for white and colored models respectively.

B. TIME DELAY

The algorithmic time delay of an enhancement system is also an important factor when a system is implemented for real-time applications. Based on the configurations given in the simulations of the earlier three chapters, $M\beta OM$ has 32 ms delay, MSKF has about 60 ms delay (including three iterations), and SANR has 18 ms delay. As a result, from the delay point of view, $M\beta OM$ and SANR are suitable for real-time applications, and MSKF is also suitable for real-time application where time delay is not very demanding.

C. TYPE OF NOISE

The three proposed methods can deal with both white noise and colored noise. They work reasonably well with stationary or slowly varying additive noise, such as interior car noise from engine, fan, wind noise and stationary wheel-to-road noise on highway/roadway, factory noise, airplane-engine noise, etc. However, in the case of highly non-stationary noise, such as interfering

voice or music, as well as the echo signal, these enhancement methods are unlikely to be effective in improving the speech quality and reducing the noise.

6.2.2 Performance Comparison

In Chapters 3, 4 and 5, we use different approaches to estimate the noise statistics in different domains, which include the frequency domain for β -order MMSE, the time domain for Kalman filtering, and the eigen-domain for signal subspace. Since these different criteria may lead to different judgement of the speech absence period given a noisy speech signal, different noise statistics may be obtained. In order to compare the three enhancement methods that operate in the different domains, we have to unify the estimation of noise statistics in the same platform for a fair comparison. Also, in order to emphasize the performance comparison of speech enhancement methods but not the performance of noise estimation, we make use of relatively stationary noises to degrade the speech signal. Consequently, the estimates of noise statistics can be computed by using the first 800 samples of each noisy speech utterance as this period is taken as a speech silence period.

In our comparison study, we use 50 phonetically balanced speech utterances from the TIMIT speech database with half from male speakers and another half from female speakers. Four types of noise from NOISEX-92 database are used for the performance evaluation; they include white noise, babble speech-like noise, F-16 flight cockpit noise and Volvo interior car noise. The sampling rate is 8000 Hz. For the proposed MSKF and SANR methods, we choose the corresponding white noise model for the case of white noise, and colored noise model for the different colored noises (such as babble, F-16 and Volvo noises).

Fig. 6.1 shows the segmental SNR performance of M β OM, MSKF and SANR with its sub-figures (a), (b), (c) and (d) corresponding to the cases of white noise, babble noise, F-16 noise and Volvo car noise, respectively. Figs. 6.2, 6.3 and 6.4 are the same as Fig. 6.1, except that the objective measures are the IS distortion, PESQ and MBSD, respectively.

As can be seen in Fig. 6.1 which shows segmental SNR performance, MSKF is almost consistently the best in the case of white noise as well as in the three different cases of colored

noise, with only a little drop at 10 dB in F-16 and Volvo noises. SANR performs better than M β OM in white and babble noises, but the performance is obviously reversed in the case of Volvo car noise. In F-16 noise, SANR yields better performance than M β OM when the input noisy segmental SNR is below 0 dB, but the latter performs better than the former when the input segmental SNR is above 0 dB and reaches the best when the input segmental SNR exceeds about 6.2 dB.

According to Fig. 6.2, M β OM has the overall best IS performance since it gives the least IS distortion for the cases of white, babble, F-16 and Volvo noise. In terms of IS distortion, MSKF performs better than SANR in the cases of white, babble and F-16 noise; however, it becomes worse than SANR in the case of Volvo noise. On the other hand, it can be seen that SANR results in more IS distortion as compared with the unprocessed noisy speech. MSKF gives improvement of IS distortion over unprocessed noisy speech in the cases of white, babble and F-16 noises, but with a loss at -10 dB in white noise case. However, it yields more IS distortion over unprocessed noisy speech in the case of Volvo noise. It should be noted that M β OM consistently contributes an improvement of IS distortion over unprocessed noisy speech in the cases of white, babble, F-16 and Volvo noise.

In terms of PESQ measure as shown in Fig. 6.3, MSKF performs the best among the three methods in the cases of white, babble, F-16 and Volvo noise, except when the input segmental SNR is high (5 dB $\sim$ 10 dB). The second best should be M β OM, whose performance is close to the MSKF's especially at high input segmental SNR. Comparing with the unprocessed noisy speech, MSKF and M β OM yield apparent improvements in PESQ score in the cases of white, babble, F-16 and Volvo noise. SANR results in improvements in PESQ score in the cases of white and F-16 noise, but gives no improvement when presented with babble or Volvo noise.

From Fig. 6.4, it can be seen that MSKF yields the best MBSD performance across the entire range of input segmental SNR under consideration in the cases of white, babble and F-16 noise. MSKF has the best MBSD measure when the input has a low segmental SNR and is only slightly poorer than M β OM in the case of Volvo noise. Basically, all three methods achieve

obvious MBSD improvement in different degrees over the unprocessed noisy speech, except for SANR in the case of babble noise.

Fig. 6.5 shows the spectrograms of a clean speech signal, a noisy speech signal contaminated by white Gaussian noise, and enhanced speech signals processed by $M\beta$ OM, MSKF and SANR, respectively. Fig. 6.6 is the same as Fig. 6.5 for the case of colored noise instead of white Gaussian noise. It can be seen that the spectrograms of the speech signals enhanced by the three enhancement methods appear much cleaner and more similar to that of the clean speech signal as compared to the unprocessed noisy speech signal. It can be observed that some weak spectral components that have been unintentionally removed with SANR are preserved very well by MSKF and $M\beta$ OM. However, $M\beta$ OM has more residual background noise as compared to the other two methods. In comparison with MSKF and $M\beta$ OM, SANT has more background noise reduced.

Informal listening tests show that the enhanced speech processed by MSKF has clearer loudness than those obtained by $M\beta$ OM and SANR, and SANR gives the perception that more background noise is removed as compared to $M\beta$ OM.

Based on the above measurements, we may conclude that the proposed MSKF method is generally the best in terms of segmental SNR, PESQ, MBSD, spectrogram and informal listening test as compared to $M\beta$ OM and SANR. However, it has the worst performance of IS measure in the presence of Volvo interior car noise. In contrast, $M\beta$ OM is the best in terms of IS distortion measure among the three methods. In the case of Volvo car noise, in terms of other three objective measures - segmental SNR, PESQ and MBSD, $M\beta$ OM loses to MSKF in low SNR and becomes a winner at high SNR.

6.3 Summary

We have compared the performances of the three speech enhancement methods that are proposed in the previous three chapters. They are $M\beta$ OM, a masking-based β -order MMSE spectral suppression method, MSKF, a masking-based subband Kalman filtering method, and SANR, a

Figure 6.1: Segmental SNRs of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and (◊) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.

Figure 6.2: IS distortion measurements of (-) input noisy speech data, processed speech data by using (*) MβOM, (+) MSKF, and (◊) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.

Figure 6.3: PESQ scores of (-) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and (◇) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.

Figure 6.4: MBS measurements of (-.) input noisy speech data, processed speech data by using (*) M β OM, (+) MSKF, and (◇) SANR with different noises: (a) white noise, (b) babble noise, (c) F-16 flight cockpit noise and (d) Volvo interior car noise.

Figure 6.5: Spectrograms of (a) clean speech (female), (b) noisy speech contaminated by white noise, processed speech by using (c) $M\beta OM$, (d) MSKF, and (e) SANR.

Figure 6.6: Spectrograms of (a) clean speech (male), (b) noisy speech contaminated by colored noise (F-16 flight cockpit noise), processed speech by using (c) M β OM, (d) MSKF, and (e) SANR.

subspace-based audible noise reduction method. All the three methods make use of the masking properties of the human auditory system, and are aimed at achieving a “perceptual solution” in reducing speech distortion while keeping the noise reduction at a certain level.

As what we may comprehend, the results obtained reveal that none of the three methods is consistently the best in all the performance measures, and they also disclose the relative strengths and weaknesses of the methods in dealing with different types of noise. Basically, MSKF and $M\beta$ OM yield better performance than SANR, especially for MSKF with the segmental SNR, PESQ and MBSD measures and $M\beta$ OM with the IS measure. From the observation of spectrograms, it is noted that MSKF yields the best enhancement effect. However, MSKF is the most computationally demanding method and has the longest time delay among the three proposed methods. It may therefore be more suitable for applications that require the highest quality for enhanced speech with less constraints on the computational complexity and time delay, e.g., speech enhancement as a front-end processing stage for a speech recognition system in a noisy environment.

As compared to MSKF, $M\beta$ OM has much lower computational complexity and a shorter time delay, and its objective measure is generally very satisfactory in terms of segmental SNR, IS, PESQ and MBSD, as compared to SANR. In particular, it is the best in terms of the IS distortion measure among the three methods. It also always leads to improvement over the unprocessed noisy speech in terms of all of measurements under consideration. Consequently, it may be suitable for use in real-time speech communication between human speakers in a noisy environment.

In fact, on the preservation of weak spectral components, the three methods have provided quite satisfactory performance, although SANR is relatively not so good in comparison. In addition, SANR has shown its effectiveness in reducing background noise, as indicated by the spectrograms as well as informal listening tests. With a compromised time delay and for a situation that demands more noise reduction, SANR may be a very suitable choice.

Chapter 7

Speech Enhancement Applications

In the preceding chapter, we compared the objective performances of the three proposed methods: masking-based β -order MMSE (M β OM) method, masking-based subband Kalman filtering (MSKF) method and subspace-based audible noise reduction (SANR) method. In this chapter, we investigate applications of the proposed speech enhancement methods for voice communication in a moving car environment and speech recognition.

7.1 Speech Enhancement for Car Mobile Communication

For voice communication in a moving car environment, the original speech is corrupted by the noise generated by the engine, fan, traffic, and wind, as well as by the transmission channel errors. Quality and intelligibility of the signals received and decoded at the receiving end may therefore be impaired. In many countries, the use of handset for communication in cars is usually restricted due to the potential risks arising from the diversion of the drivers' attention. Hands-free devices are in general allowed in order to overcome this problem. The convenience of hands-free phone usage has prompted many studies to enhance the quality of the received speech. The speech quality available in a moving car environment, however, is severely degraded when a hands-free telephone is used due to the presence of the ambient noise. In a communication system, speech signals are typically coded using a speech model with characteristics tuned to clean speech signals, and it is therefore normally not suitable for the noisy signals. Consequently,

the presence of acoustic background noise degrades the performance of a speech codec far more than would be expected from an input noisy speech signal alone without coding. In fact, a low SNR at the input can lead to errors in the estimation of model parameters used in a speech coder. In addition, low quality speech is stressful and fatiguing over long listening durations, and is often accompanied by reduced speech intelligibility, especially while one is driving a car. Similarly, if a mobile communication system is equipped with a speech recognition system which is used for automatic dialing, the recognition accuracy of such a system deteriorates in the presence of noise, since the noisy input signal is unlikely to obey the statistical model of the clean speech signal used by the recognizer. Therefore it is of interest to investigate the effectiveness of a speech enhancement approach in improving the performance of a voice communication system in a moving car environment.

There have been many speech enhancement approaches applied to mobile voice communication in a moving car environment for improving both communication comfort and recognition rates of voice-controlled systems [75, 87]. As we only focus on a unique type of noise, i.e., running car noise, it may be possible to further improve the performance of a speech enhancement system provided that we know the generalizable characteristics of the noise spectrum. Since $M\beta OM$ shows a better IS performance as compared to the MSKF and SANR methods, and it also has a satisfactory performance in the case of the Volvo interior car noise as well as not too high a computational complexity and algorithmic time delay. In this section, $M\beta OM$ is chosen to be applied for mobile communication in a moving car. In order to achieve more effective preservation of weak speech spectral components and reduction of perceptible noise, we modify $M\beta OM$ to improve its effectiveness in a car noise environment.

7.1.1 Modified Adaptation of β Value

In a running car, most of the noise come from the engine, the revolving mechanical parts, the air flow from the air ducts, wind, road surface (tire) noise, and the passing vehicles. The noise from engine and mechanical parts of the car is a relatively slowly varying process, while the noise from passing cars is usually nonstationary and transient. Most researchers consider car noise in

general as a predominantly low frequency noise with most of its energy falling below 500 Hz. Fig. 7.1 shows the power spectral density of the noise signal in a car which is moving at various speeds. It can be seen from the figure that the spectral energy distribution of the interior car noise drops rapidly from its peak at about 50 ~ 60 Hz as frequency increases.

In comparison with an ordinary type of noise, the low frequency energy of interior car noise usually appears to be extremely strong in contrast with the high frequency energy of the noise which is relatively very weak in general. With such kind of extreme characteristics, it is logical to adjust the adaptation of the β value of the original M β OM method, so as to provide a more accurate estimation of the speech signal in consideration of the characteristics of interior car noise. As explained in the earlier chapter, a high β value in the M β OM method leads to more preservation of speech spectral components and less reduction of the noise components. It is therefore reasonable to choose a small part of the dynamic range of β value to be dependent on the frequency when the shape of a stationary noise spectrum is already known. Based on the spectral shape of the interior car noise, we proposed a function of frequency to be as follows

$$\Delta(f) = 0.56 \left(\left| \frac{f-50}{4000} \right|^\zeta + \left| \frac{f-60}{4000} \right|^\zeta - 0.21 \right), \quad 0 \leq f \leq 4000 \text{ (Hz)} \quad (7.1)$$

where f is the frequency, and ζ is an adjustment factor which is chosen empirically to be in the range of [0.15, 0.26]. The curves of the function $\Delta(f)$ with two different ζ values are shown in Fig. 7.2. In our simulations, we set ζ to 0.23.

Based on the β adaptation equation of M β OM - Eq. (3.44), the β adaptation of the modified M β OM is given by

$$\begin{aligned} \tilde{\beta}(l, k) &= 1.42\Delta(f) + 0.121\Xi(l) + 0.981\Theta_f(l, k) + 0.187 \max[\Xi(l) + 6.7, \quad 0]\Theta_f(l, k) \\ &= 0.7952 \left(\left| \frac{k}{K} - 0.0125 \right|^\zeta + \left| \frac{k}{K} - 0.015 \right|^\zeta - 0.21 \right) + \\ &\quad 0.121\Xi(l) + 0.981\Theta_f(l, k) + 0.187 \max[\Xi(l) + 6.7, \quad 0]\Theta_f(l, k) \end{aligned} \quad (7.2)$$

where K denotes the number of frequency bins and k the frequency bin number which is related to f by $k = \frac{K}{4000}f$. The dynamic range of β value is the same as that of the original M β OM, i.e.,

$$\hat{\beta}(l, k) = \min \{ \max\{\tilde{\beta}(l, k), \quad 0.001\}, \quad 4 \}. \quad (7.3)$$

(a)

(b)

Figure 7.1: Power spectrum measure in a car

Figure 7.2: Curve of $\Delta(f)$ function

7.1.2 Performance Evaluation

The performance evaluation of the proposed enhancement algorithm is based on spectrogram analysis, objective assessment using the IS distortion and the MBSD measure, and subjective listening tests. We generate noisy signals by artificially adding real single-channel car noise data to clean speech signals with different SNRs. Various car noise types are used in our performance evaluation. These include the noise recorded with speeds at 50, 70, 90, 110, 115 and 120 km/h under different car running environments including high-way, road-way, rainy-road, and opened/closed window. A total of 30 different utterances from the TIMIT database are used in our evaluation. Half of the utterances are male and the other half female. The effectiveness of the proposed enhancement algorithm is evaluated for the 8kHz sampling rate case with the frame size of 256 samples, which are Hanning windowed with 75% overlap between adjacent frames. We use 256 point FFT, so $K = 128$.

Fig. 7.5 shows the IS distortion and MBSD from the enhanced results by using different

estimators with various input segmental SNRs with speed at 70 km/h in the case of car window wound up.

Figs. 7.6 and 7.7 are the same as Fig. 7.5 except that the car noises are obtained at the speeds of 90 and 70 km/h respectively.

In addition, we also investigated the car noise in the window opened (windy) scenario. Fig. 7.8 shows the IS distortion and MBSD from the enhanced results by using different estimators for various input segmental SNRs. Figs. 7.9 and 7.10 are the same as Fig. 7.8 except that the car noises are obtained at the speeds of 90 and 70 km/h respectively.

Figs. 7.3 and 7.4 show the spectrograms, where the speech signals are corrupted by 115km/h highway car-window-close noise, obtained by the proposed speech enhancement method in comparison with the LSA estimator.

The above performance results indicate that the modified $M\beta$ OM always outperforms other methods under consideration in terms of the IS distortion measure, MBSD measure and spectrogram inspection. Informal listening tests also confirm the advantage of our modified $M\beta$ OM applied in different car noise conditions.

7.2 Speech Enhancement for Speech Recognition

One of the major challenges in speech recognition research is to develop the solution for robust recognition of noisy speech [119]. The degradation of the speech recognition performance is known to be due to the mismatch between the trained templates (or models) and the templates (or models) under test. One of the solutions that may be considered is to obtain the reference models by using the noisy speech signals, instead of the clean speech signals, for training. However, the approach is very unreliable and almost impractical when the noise environment or SNR varies. Many past simulation studies [12, 109] indicate that the success rate of this approach is not improved when it is being tested with the different type of noise or different SNR used for training as compared with the case where clean speech is used for training. In order to solve the noise degradation problem, many research papers have reported techniques which include

Figure 7.3: Speech waveforms and spectrograms: (a) Clean speech (b) Noisy speech (90 km/h car noise with window open) (c) LSA estimated speech signal (d) Modified $M\beta OM$ estimated speech signal.

Figure 7.4: Speech waveforms and spectrograms: (a) Clean speech (b) Noisy speech (115km/h car noise with window close) (c) LSA estimated speech signal (d) Modified M β OM estimated speech signal.

(a)

(b)

Figure 7.5: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window close at speech 70 km/h.

(a)

(b)

Figure 7.6: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window close at speech 90 km/h.

(a)

(b)

Figure 7.7: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window close at speech 115 km/h.

(a)

(b)

Figure 7.8: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window open at speech 70 km/h.

(a)

(b)

Figure 7.9: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified MβOM (star) with car noise at highway with window open at speech 90 km/h.

(a)

(b)

Figure 7.10: Objective performance measures in terms of (a) IS distortion, and (b) MBSD measure, with various estimators: LSA (circle), E-M MMSE (diamond), Wiener filtering (dashed-dot), and modified M β OM (star) with car noise at highway with window open at speech 110 km/h.

speech enhancement as pre-processing [50, 58], speech feature enhancement [56, 61], feature compensation [30] and model-based feature enhancement for speech recognition [5, 108].

Research in speech recognition has been active for a long time and a large number of algorithms and performance results are available [19, 66, 94]. Usually, speech recognition consists of a training phase and a recognition phase. Both phases involve extraction of features which represent the unique set of characteristics of a speech segment and are thus used to represent the segment. The selection of features affects directly the performance of a speech recognition system. So far, features such as 12 Mel frequency cepstral coefficients (MFCC) and energy, their differences (delta coefficients) and double-differences (acceleration coefficients) are typically adopted in many speech recognition systems. In the training phase, the system learns the reference patterns representing the different speech sounds that constitute the vocabulary of the application. Each reference is trained using spoken utterances in the training database and stored either in the form of templates obtained by some averaging methods or models that characterize the statistical properties of the patterns. In the recognition phase, an unknown speech signal is to be recognized by comparing with the trained models. Fig. 7.11 illustrates the two phases of a typical speech recognition system.

Figure 7.11: The training phase and recognition phase of a typical speech recognition system.

The problem of speech recognition can roughly be considered as the decoding of the information of a speech signal and its transcription into a set of characters. These characters can be

subsequently used to perform various tasks. The difficulties of speech recognition system may include large intra-class variations, sometimes small inter-class variations, the detection of end-point for continuous speech, large vocabulary, complexity of languages, and the environmental conditions such as various reverberations, time-varying channel characteristics and background noise.

There have been many successful applications [113] such as voice control roboticized functions [88], speech-to-text transcription [57], intelligent Internet browsing via voice [29] and voice-activated telephone system. Telephony-based speech recognition systems are gaining more popularity these days. Typical applications include automatic call centers, remote voice inquiries and commercial transactions. Currently, most commercial general-purpose speech recognizers can provide acceptable recognition accuracy for relatively clean speech. However, the performance of such recognition systems degrades significantly when they are applied to real situations where there are various noisy interferences. Such problems may be more serious for the telephony-based recognition system, where the remote users may use speaker-phone or mobile phone in some noisy environments.

In this section, we study the telephony-based speech recognition system working in different background noises. According to the performance of MSKF as described in the last chapter, MSKF is generally the best among the three proposed masking-based speech enhancement methods in terms of segmental SNR, PESQ, MBSQ, spectrogram inspection as well as informal listening test. Although the computational complexity and the time delay of MSKF method are higher than the other two methods, i.e., M β OM and SANR, the requirement of the computational complexity and the time delay of speech enhancement for a speech recognition system is not demanding as compared with speech enhancement for phone communication. Therefore, we choose our proposed MSKF speech enhancement method for preprocessing in a telephony speech recognition system which is realized by using Nuance 8.0 commercial software. We show the effectiveness of MSKF enhancement method in improving the speech recognition rate of telephony-based name recognition.

7.2.1 Introduction of Speech Recognition System

There are many commercial speech recognition engines for telephony applications. The major vendors include SpeechWorks, Nuance, Locus, Philips and Vocalis. SpeechWorks develops speech recognition engine technology for over-the-telephone customer service and transactions. Nuance offers a continuous speech recognizer for large vocabulary applications, including stock quotes. Locus Speech Corporation has recently announced a speaker-independent, flexible vocabulary telephony speech engine in English and French. Philips Speech Processing offers leading edge speech recognition, including natural dialogue systems for telephony, consumer products and the IT industry. Vocalis focuses on speech for computer telephony in Europe.

In this chapter, we study the use of Nuance speech recognition engine for our speech recognition application to evaluate the performance of our proposed MSKF speech enhancement method. Similar to common up-to-date speech recognition systems, Nuance 8.0 system also completes the recognition through end-point detection (or segmentation), feature extraction, vector quantization, recognition and understanding (according to grammar). In the feature extraction phase, the Nuance system examines the MFCC-based (Mel-Frequency Cepstral Coefficient) feature set, a vector, in the frequency band at range of 300 Hz to 3.4 kHz. This vector is then transformed into a new vector well suited for speech recognition processing. In the decoding phase, the Nuance system uses hidden Markov models (HMMs) as the acoustic models for mapping the utterance signal to phoneme sequences. HMMs are a powerful modeling technique for discrete state processes [62, 65]. The basic principle of HMM is that the observation sequence is generated by a system that can exist in one of a finite number of states. At each time-step, the system makes a transition from the current state to another state, and emits an observable quantity according to a state-specific probability distribution. HMM provides a detailed spectral and temporal representation of the speech signal. These models automatically learn from data and have been optimized for telephone-quality speech signal. Because of the time-varying nature of the speech signal, it is important for a recognition system to have a mechanism to deal with time. In HMMs, a stationary Markov chain is used as this mechanism. The basic premise is that the probability

of transitioning from one state to another depends only on the two states and not on any prior transitions.

Nuance uses Markov models at many levels of the system. Firstly a speech signal is broken down into words, and they may be referred to as the sentence level. At this level, a grammar database is used. For name recognition, it could be referred to as the name-grammar. Secondly, words are segmented into phones which are linked to a dictionary at the phone level. Thirdly, the phonetic units are further broken down into start, middle and end states. This is the lowest level that links to the observed signal directly. Each of these sequences is modeled with a Markov chain. At the lowest level, states are allowed to transit to themselves. The self-loop allows HMMs to model the duration of a sound, which is very important because of the high variability in speaking rate for different talkers.

In speech recognition of names, an HMM is constructed for every set of speech data. The system then computes the probability that the data were generated by each HMM. The recognition result is the name that corresponds to the HMM that generated the observed speech data with the highest probability. Though this approach seems straightforward, it is computationally demanding to explicitly evaluate each possible sentence when the number of the candidate sentences is not small. To overcome this problem, the Nuance recognition system makes use of the hierarchical HMM structure, i.e., sentence HMMs are a concatenation of word HMMs; word HMMs are a concatenation of phonetic unit HMMs, and so forth.

During the recognition stage, the computation can be shared across hypothesized sentences. However, even the sharing of computation is not sufficiently efficient for real-time recognition, and pruning which deletes unlikely hypotheses is required. The search method used by the Nuance system is the well known Viterbi algorithm [23]. Although this pruning search strategy is very efficient and gives the Nuance system the capability to recognize vocabularies of more than 10,000 words uttered in almost unlimited number of possible ways, it is at the expense of increased recognition errors.

7.2.2 Speech Enhancement for Speech Recognition

We use the Nuance 8.0 speech recognition system which works in a batch mode and is set specifically for telephony recognition to verify the effects of speech enhancement on the recognition rate. The recognition task consisted of recognizing names (in English) contained in a database, in a speaker-independent mode. The baseline was established by selecting 650 out of 739 names uttered by each speaker, such that the recognition rate of clean speech was achieved at 90%, where there were a total of 739 candidate names in the database, which were spoken by each of five speakers (three females and two males), or a total of 3250 from 3695 individual utterances. It means the upper bound of recognition rate is 90%, and the statistical lower bound is $\frac{1}{739} \simeq 0.1353\%$. The speech samples were recorded using the handset of a fixed-line telephone, in a relatively quiet office environment. The SNRs of the speech signals in the database are above 35 dB. In view of this relatively clean name-database, the inherent noise may be neglected. The various background noises were then digitally added to the audio files in the database to produce various SNR levels. In our simulation, we chose the input segmental SNR of -5, 5, 15 and 25 dB¹.

Two types of noise from the NOISEX-92 database, namely white noise and babble mass-speech-like noise, were selected in our simulation study. As we were concerned with telephony speech recognition, the noises had to be processed by a bandpass filter (300-3400 Hz) before they were added to the speech samples from the speech database. Fig. 7.12 illustrates the performance evaluation flowchart of the speech enhancement methods in our speech recognition experiments. We selected the MSKF method for the pre-processing stage of the speech recognition system. Same as what has been mentioned in the previous chapter, the MSKF method with white noise model is applied to the white noise case, and its colored noise model is adopted for colored noise reduction. In the experiments, we obtained the recognition rates in both ideal and real cases².

¹Since the recorded name speech waveforms have quite a long speech absence period in a waveform, we measure the input segmental SNR by selecting the segments where the energy is higher than an empirical small threshold. Then we obtained the input segmental SNR by averaging the SNR of all of the selected segments in a whole speech sample.

²As described in chapter 4, the ideal case means that the AR parameters and the masking threshold are directly

Figure 7.12: Block diagram of performance evaluation for speech recognition

Figs. 7.13 and 7.14 show the recognition success rates for the cases of white noise and colored (babble) noise, respectively. For the white noise case, degradation in recognition accuracy due to the noise is very severe, especially when the input segmental SNR is low. Through enhancement processing by using the proposed MSKF method, it can be observed that a significant improvement in the recognition rate is achieved, although there is still an obvious performance gap between both real and ideal cases. The recognition accuracy is improved from 5.53% to 38.95% after enhanced by MSKF with real case when the input segmental SNR is -5 dB.

For the (babble) colored noise case, the success rate drops even more severely due to noise degradation as compared with the white noise case. The recognition accuracy is raised from 1.53% to 33.89% after enhancement by MSKF for the input segmental SNR of -5 dB. However, as shown in Fig. 7.14, the degradation in recognition rate due to noise is less severe than that in the white noise case when input segmental SNR is high. It can also be seen that with increasing input segmental SNR, in contrast with the white noise case, the recognition rate improvement due to the use of enhancement is not obvious and there is almost no improvement arising from computed by using the clean speech signal; the real case means that only noisy speech is available for realistic implementation and the enhanced speech is estimated through three iterations of MSKF method.

Figure 7.13: Speech recognition success rate in the presence of white noise.

Figure 7.14: Speech recognition success rate in the presence of colored (babble) noise.

the enhancement processing when the segmental SNR reaches 25 dB.

In our application system the speech signal is pre-processed via MSKF enhancement and the enhanced speech is then passed to a separate feature extraction routine. As we use the commercial software to implement the speech recognition part including feature extraction and information decoding, we cannot combine the enhancement method and feature extraction parts together which is more like the conventional speech feature enhancement. It is worth mentioning that the robustness and computational complexity may be potentially improved if the two parts can be combined together.

7.3 Summary

In this chapter, two proposed speech enhancement methods have been exploited in two particular applications: voice communication in a moving car environment and a speech recognition system with noisy speech input.

By using the frequency characteristics of real-life car noises, our modified $M\beta OM$ algorithm is confirmed through a large number of experiments to be significantly effective in car mobile communication application. The simulation results prove that the modified $M\beta OM$ method outperforms many existing methods in car mobile application, especially for the case of weak spectral components of speech signal corrupted by car noise.

The effectiveness of speech enhancement is studied for a speech recognition system whose goal is to automatically recognize the names spoken by a speaker via signal processing in a noisy environment. The Nuance speech recognition engine is chosen for the speech recognition system and the MSKF speech enhancement method is applied as the front-end processing stage, and the performance of the enhancement method is evaluated through a real-life recorded name database in terms of the recognition success rate as compared to the case of unprocessed noisy samples. It is observed that the recognition success rate is significantly raised after the noisy speech is enhanced by MSKF in almost the entire SNR range under consideration, except for the input segmental SNR of more than 25 dB for the babble noise case. It is shown that the MSKF method

is effective as a front-end processing stage in a speech recognition system.

Chapter 8

Conclusion and Recommendations

8.1 Conclusion

In this thesis, various techniques for single channel speech enhancement have been investigated. The main objective of speech enhancement is to suppress the noise components while keeping the distortion of the speech signals to a minimum level. It is observed that artifacts always occur while the noise level is reduced. How to optimally achieve a trade-off between the amount of artifacts and the level of reduction of the noise and how to shape the artifacts into an inaudible form have been the main focus of most speech enhancement methods. Exploitation of the psychoacoustic model in a speech enhancement system is a dominant idea in seeking a perceptually optimal trade-off between speech distortion and noise reduction. Therefore, the main purpose of masking-based speech enhancement methods is to minimize the distortion of speech components while constraining the audible noise to a certain level.

β -order STSA MMSE estimation has been introduced and the characteristics of the β -order STSA MMSE estimator has also been examined. As compared to conventional STSA MMSE and LSA, the β -order MMSE estimator has more room in achieving a more optimal estimation of speech components given a noisy signal. Comparing to the generalized spectral subtraction method, β -order MMSE involves more stringent mathematical derivation and is based on an appropriate statistical model assumption. The relationship between the value of β and the spectral

amplitude gain function of the MMSE method is analyzed. Since the implementation of the β -order MMSE involves the computation of hypergeometric function and gamma function which are of high computational complexity, for practical implementation, an appropriate function table look up to reduce the computational complexity is proposed and at the same time the amount of memory used for the tables is saved. Following that, adaptive β -order MMSE algorithms have been developed and shown to result in even more significant performance improvement, especially for the case of time-frequency adaptation of β value which leads to more preservation of weak speech spectral components. Its truly significant improved performance, as compared to many algorithms, in terms of both objective and subjective assessments for a large number of speech data, many noise types and various SNR levels, has been amply demonstrated. Moreover, it is further developed into a masking-based β -order STSA-MMSE ($M\beta OM$) speech enhancement method, in which the adaptation of the value of β is based on the perceptual properties of the human auditory system. Through computer simulations, our proposed masking-based β -order MMSE method has been shown to outperform many other speech enhancement methods significantly and has achieved a good tradeoff between minimizing speech distortion and maximizing residual noise reduction, especially for the case of weak spectral components of a speech signal corrupted by noise. Among the proposed modifications of adaptive β -order MMSE, the introduction of speech-presence uncertainty leads to more reduction of noise and the exploitation of masking properties helps to preserve weak speech spectral components through reducing only the perceptible noise.

In this thesis, a masking-based Kalman filtering (MSKF) method is developed, which involves the use of masking properties, subband decomposition and Kalman filtering techniques. Based on the standard Kalman filtering concept and conventional AR parameter estimation, the proposed Kalman filtering method for speech enhancement is implemented. A pseudo-QMF is used to implement subband decomposition and a relation between the masking threshold and the noise variance estimate used in Kalman filtering is established. Kalman filtering speech enhancement in the subband domain not only has a much lower computational complexity, but also a higher

SNR improvement than when it is implemented in the fullband domain. The incorporation of masking properties in the proposed subband Kalman filtering method is found to have significant performance improvement over the conventional subband Kalman filtering scheme. Through simulations, the proposed scheme has been examined by studying the contributions from different functions including Kalman filtering, subband decomposition and perceptual effect based on masking threshold. The optimal configuration is investigated, which leads to better enhancement results as compared to the fullband and the normal subband Kalman filtering methods. Through a large number of simulations, it is shown that the proposed masking-based Kalman filtering method outperforms many existing well known methods in terms of objective as well as subjective measures.

In addition to the first two contributions, a subspace-based audible noise reduction (SANR) scheme has been also proposed. Firstly, an invertible frequency eigendomain transformation is derived, which correctly reflects the true quantity of the eigenvalues transformed from the corresponding PSD or masking threshold. By using the transformation, masking properties of the human auditory system can be incorporated into subspace speech enhancement methods. Based on the knowledge of the human auditory system, the suppression of the audible noise quantity is focused in the eigendomain, and the level of reduction for inaudible noise is relaxed so that some weak components of the underlying speech signals are effectively preserved and the distortion of the speech signals are therefore reduced. The eigen-decomposition of the estimated speech autocorrelation matrix with the assumption of white noise is derived. The masking properties of the human auditory system are incorporated into the signal subspace technique through the definition of the residual audible eigen-noise. The gain of eigen-filter is proposed to be regulated by an audible noise reduction criterion for the white noise model. Subsequently, an audible noise reduction scheme is developed based on the signal subspace technique and the implementation of our proposed scheme is outlined. The scheme is further extended to the colored noise model. Simulation results show the superiority of our proposed scheme over other existing subspace methods in terms of segmental SNR, spectrogram, PESQ and informal listening tests.

Moreover the three proposed masking-based enhancement methods are compared in terms of objective and subjective measures through a large number of simulations, and discussed their individual strengths and weaknesses.

Finally, the application of speech enhancement for processing noisy speech contaminated by interior car noise for mobile voice communication are investigated. A modified M/β OM scheme is proposed to achieve car noise reduction for various types of moving car noise arising from different speeds and conditions. Furthermore a speech recognition system is described and the effectiveness of MSKF speech enhancement method for the speech recognition system is studied. It was demonstrated through computer simulations using a large number of real-life recorded data that MSKF led to a significant improvement in the robustness of the recognition system.

8.2 Recommendations for Further Research

Based on the methods proposed in this thesis, we summarize our suggestions for future research in speech enhancement as follows:

Adaptive β -order MMSE provides some significant improvement in terms of objective and subjective performance measures. However, the adaptation of the β value depends on less than perfect analysis and empirical adjustments. In future work, we suggest to derive a more optimal adaptation of β by taking into account the correlations among the spectral components. It may be considered to establish a mathematical model to incorporate masking properties into β -order MMSE estimation instead of the somewhat empirical approach adopted in our research.

Both Wiener filter and Kalman filter can be made to optimally achieve the same objective, i.e., the linear MMSE estimation of a wide-sense stationary signal from a noisy observation. Kalman filter has an additional advantage in that it can be used for nonstationary signals. With the introduction of AR speech production model, Kalman filtering can efficiently exploit speech production characteristics, and therefore outperforms Wiener filtering. Although various attempts have been made to incorporate the Kalman filtering technique into speech enhancement systems, it is still not clear what the best way of applying Kalman filtering to speech enhancement is. It

implies that the existing Kalman filtering based algorithms have some deficiencies that deserve further improvements. The proposed masking-based Kalman filtering method uses uniform subbands to map to non-uniform critical bands, where the error varies depending on the number of subbands. Though the uniform subband scheme provides a way to achieve a reasonable incorporation of masking properties into a Kalman filtering speech enhancement scheme and it also reduces the computational complexity, the slightly mismatched error¹ arising from the use of uniform subbands can be avoided when we adopt non-uniform subbands that are the same as the critical bands. Thus, we suggest to investigate the effectiveness of the masking-based Kalman filtering method using non-uniform subbands in place of uniform subbands. It may be consider also to develop a simplified non-uniform subband scheme in order to reduce the implementation complexity. The principle of linear Kalman filtering is based on the orthogonality between the innovation process and the past observation. This ideal orthogonality condition is often violated due to the deviations caused by the model used in a real signal. This simply implies that the Kalman filter used in our scheme is not always optimal. It might therefore be fruitful to investigate the masking-based Kalman filtering speech enhancement method with a modified Kalman filtering method proposed in [97].

The proposed subspace-based audible noise reduction scheme requires a large amount of computations for the invertible frequency eigen-domain transformation. In future work, we suggest to design a simplified algorithm to reduce its computational complexity. It is also possible to combine eigen-decomposition and masking threshold computation to reduce the overall computational complexity of the system.

Speech recognition performance is known to degrade dramatically when the speech signals are contaminated by noise. This is due to the mismatch between the trained models and the features under test. In the use of masking-based speech enhancement to improve the robustness of a speech recognition system, the overall computational effort needed by the system is determined to a certain extent by the pre-processing transformation from the time domain to the domain

¹Since we know that the masking threshold is computed according to the critical band, but not uniform band; i.e., the thresholds of the frequency bins falling within the same critical band are considered to be same value.

which suits the recognition process. We may therefore consider combining both feature extraction used in speech recognition and speech enhancement processing in the same domain, in order to design a more coherent and robust feature extraction scheme to achieve better recognition performances.

Author's Publications

Journal Papers

1. C.H. You, S.N. Koh and S. Rahardja, " β -Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, July 2005.
2. C.H. You, S.N. Koh, and S. Rahardja, "Masking-Based β -Order MMSE Speech Enhancement", *Speech Communication*, Vol. 48, pp. 57-70, Jan. 2006.
3. C.H. You, S.N. Koh, and S. Rahardja, "An Invertible Frequency Eigendomain Transformation for Masking-Based Subspace Speech Enhancement", *IEEE Signal Processing Letter*, Volume 12, No. 5, pp. 461-464, May 2005.
4. C.H. You, S. Rahardja, and S.N. Koh, "Signal Subspace Speech Enhancement Using Masking Properties", Submitted to Journal, 2006.
5. C.H. You, S.N. Koh, and S. Rahardja, "Subband Kalman Filtering Incorporating Masking Properties for Noisy Speech Signal", Submitted to Journal, 2006.

Conference Papers

1. C.H. You, S. Rahardja, and S.N. Koh, "Perceptual Kalman Filtering Speech Enhancement," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-06, Vol. 1, pp. 461-464, France, May 14-19, 2006.

2. C. H. You, S. N. Koh, S. Rahardja, L. Shue, F. Sattar, "Experiment of Masking-Based beta-Order MMSE Speech Enhancement in a Car Noise Environment," *Proc. Int. Conf. Information, Communication and Signal Processing*, ICICS, pp. 715-718, Thailand, Dec. 2005.
3. C.H. You, S.N. Koh, and S. Rahardja, "Subspace Speech Enhancement for Audible Noise Reduction", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-05, Vol. 1, pp. 145-148, USA, Mar. 2005 ².
4. C.H. You, S.N. Koh, and S. Rahardja, "An MMSE Speech Enhancement Approach Incorporating Masking Properties", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-04, Vol. 1, pp. 725-728, Canada, May 2004.
5. C.H. You, S.N. Koh and S. Rahardja, "Kalman Filtering Speech Enhancement Incorporating Masking Properties for Mobile Communication in a Car Environment," *Proc. IEEE International Conference on Multimedia and Expo*, ICME'2004, pp. 1343-1346, Jun. 2004
6. C.H. You, S.N. Koh, and S. Rahardja, "Adaptive β -Order MMSE Speech Enhancement Application for Mobile Communication in a Car Environment", *Proc. Int. Conf. Information, Communication and Signal Processing*, ICICS, Singapore, Dec. 2003.
7. C.H. You, S.N. Koh, and S. Rahardja, "Adaptive β -Order MMSE Estimation for Speech Enhancement", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-03, Vol. 1, pp. 852-855, Hong Kong, 2003.

²This paper is rated among the top in the Speech Signal Processing category of ICASSP-2005 by the IEEE Signal Processing Society, and invited to submit to IEEE journal.

Bibliography

- [1] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, Feb. 2001.
- [2] A. Varga and H. 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*, Vol.12, No. 3, pp. 247-251, Jul. 1993.
- [3] B. De Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. on Signal Processing*, vol. 41, no. 9, pp. 2826-2838, Sept. 1993.
- [4] B. Delgutte and N.Y.S. Kiang (1984), "Speech coding in the auditory nerve: I. Vowel-like sounds", *Journal of Acoustical Society of America*, JASA Vol. 75, pp. 866-878, 1984.
- [5] C. Couvreur and H. Van hamme, "Model-based feature enhancement for noisy speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 3, pp. 1719-1722, June 2000.
- [6] C.H. You, S.N. Koh and S. Rahardja, " β -Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, July 2005.
- [7] C.H. You, S.N. Koh and S. Rahardja, "Kalman Filtering Speech Enhancement Incorporating Masking Properties for Mobile Communication in a Car Environment," *Proc. IEEE International Conference on Multimedia and Expo, ICME'2004*, pp. 1343-1346, Jun. 2004

- [8] C.H. You, S.N. Koh, and S. Rahardja, "Adaptive β -Order MMSE Estimation for Speech Enhancement", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-03, Vol. 1, pp. 852-855, 2003.
- [9] C.H. You, S.N. Koh, and S. Rahardja, "An Invertible Frequency Eigendomain Transformation for Masking-Based Subspace Speech Enhancement", *IEEE Signal Processing Letter*, Volume 12, No. 5, pp. 461-464, May 2005.
- [10] C.H. You, S.N. Koh, and S. Rahardja, "Masking-Based β -Order MMSE Speech Enhancement", *Speech Communication*, Vol. 48, pp. 57-70, Jan. 2006.
- [11] C.H. You, S.N. Koh, and S. Rahardja, "Subspace Speech Enhancement for Audible Noise Reduction", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-05, Accepted, Mar. 2005.
- [12] C. Meyer and G. Rose, "Improved Noise Robustness by Corrective and Rival Training," *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 293-296, May 2001.
- [13] D. Malah, R.V. Cox, and A.J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-99, pp.789-792, 1999.
- [14] D.E. Tsoukalas, M. Paraskevas and J.N. Mourjopoulos, "Speech Enhancement Using Psychoacoustic Criteria," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-93., Vol. 2, pp. 359-362, 1993.
- [15] D.E. Tsoukalas, J.N. Mourjopoulos and G. Kokkinakis, "Speech Enhancement Based on Audible Noise Suppression," *IEEE Trans. Speech and Audio Processing*, Vol. 5, No.6, pp. 497-514, Nov. 1997.
- [16] D. Sinha and A.H. Tewfik, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelets," *IEEE Trans. Signal Processing*, Vol. 41, No. 12, pp. 3463-3479, Dec. 1993.

- [17] E. Zwicker, "Subdivisions of the audible frequency range into critical bands," *Journal of the Acoustical Society of America*, No. 33, pp. 248, 1961.
- [18] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Berlin, Springer-Verlag, 1990.
- [19] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. IEEE*, Vol. 64, No. 4, pp. 532-556, 1976.
- [20] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 700-708, Nov. 2003.
- [21] F. Jabloun and B. Champagne, "On the Use of Masking Properties of the Human Ear in the Signal Subspace Speech Enhancement Approach," in *proc. of Int. Workshop on Acoustic Echo and Noise Control, IWAENC*, Darmstadt, pp. 199202, Germany, Sept. 2001.
- [22] F. Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, ICASSP-02*, Vol. 1, pp. 569-572, May 2002.
- [23] G. David Forney, Jr. "The Viterbi algorithm," *Proc. of the IEEE*, Vol. 61, No. 3, pp. 268-278. March, 1973.
- [24] G.U. Yule, "On a method of investigating periodicities in disturbed series with special references to Wolfers sunspot numbers," *Philos. Trans. R. Soc. London, ser. A* 226, pp. 267C298, 1927.
- [25] G.T. WALKER, "On periodicity in series of related terms," *Proc. Roy. Soc. London Ser. A* 131, pp. 518C532, 1931.
- [26] G.A. Miller, *Language and Communication*, New York: McGraw-Hill, p. 61, 1963.

- [27] H.G. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-95, Vol. 1, pp. 153-156, 1995.
- [28] H. Fletcher, and W. A. Munson, "Loudness, its definition, measurement and calculation," *Journal of the Acoustical Society of America*, No. 5, pp. 82C108, 1933.
- [29] H. Gu, J. Li, B. Walters, and E. Chang, "Spoken Query for Web Search and Navigation," *Poster Proc. 10th Int. World-Wide Web Conf.*, May 2001.
- [30] H. K. Kim, R. C. Rose and H. G. Kang, "Acoustic Feature Compensation Based on Decomposition of Speech and Noise for ASR in Noisy Environments," *Proc. of EuroSpeech*, pp. 421-424, 2001.
- [31] H. Lev-Ari and Y. Ephraim "Extension of the Signal Subspace Speech Enhancement Approach to Colored Noise," *IEEE Signal Processing Letters*, Vol. 10, No. 4, Apr. 2003.
- [32] H. Sameti, H. Sheikhzadeh, L. Deng and R. L. Brennan "HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 5, Sep. 1998.
- [33] H. Xu, Z.H. Tan, P. Dalsgaard and B. Lindberg, Spectral Subtraction with Full-Wave Rectification and Likelihood Controlled Instantaneous Noise Estimation for Robust Speech Recognition, *Proc. of International Conference on Spoken Language Processing*, ICSLP (INTERSPEECH), Korea, Oct. 2004.
- [34] I. B. Thomas and R.J. Niederjohn, "Enhancement of Speech Intelligibility at high noise levels by Filtering and Clipping," *J. Aud. Eng. Soc.*, Vol. 16, pp. 412-415, 1968.
- [35] I. B. Thomas, "The Second Formant and Speech Intelligibility," in *Proc. Nat. Electronics Conf.*, Vol. 23, pp. 544-548, 1967.
- [36] I.Y. Soon and S.N. Koh, "Speech Enhancement Using 2-D Fourier Transform," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 6, pp. 717-724, Nov. 2003.

- [37] I.Y. Soon, S.N. Koh, and C.K. Yeo, "Improved Noise Suppression Filter Using Self-Adaptive Estimator of Probability of Speech Absence," *Signal Processing*, Vol.75, No.2, pp 151-159, June 1999.
- [38] I. Y. Soon, S. N. Koh and C. K. Yeo, "Wavelet for speech denoising," *Proc. of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications*, TENCON-97, Vol. 2, pp. 479-482, Dec. 1997.
- [39] I. Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, Nov. 2001.
- [40] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.
- [41] ISO/IEC, *Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s*, Part 3: Audio, International Standard, ISO/IEC 11172-3: 1993.
- [42] J. Benesty, A. Gilloire, and Y. Grenier, "A frequency-domain stereophonic acoustic echo canceler exploiting the coherence between the channels," *Journal of Acoustical Society of America*, vol. 106, pp. L30- L35, Sept. 1999.
- [43] J. Jensen and J.H.L. Hansen, "Speech Enhancement Based on a Constrained Sinusoidal Model," *Inter. Conf. Spoken Language Processing*, Vol. 2, pp. 214-217, Oct. 2000.
- [44] J.M. Kates, "Speech enhancement based on sinusoidal model," *J. Speech Hearing Research*, Vol. 37, No. 2, pp. 449-464, 1994.
- [45] J.S. Garofolo, "Getting Started with the DARPA TIMIT CD-ROM: an Acoustic Phonetic Continuous Speech Database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, (prototype as of December 1988).
- [46] J.S. Lim and A.V. Oppenheim, "All-Pole Modelling of Degraded Speech," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. 26, pp. 197-210, Jun. 1978.

- [47] J.S. Lim, "Evaluation of a Correlation Subtraction Method Enhancing Speech Degraded by Additive White Noise," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, No. 5, pp. 471-472, Oct. 1978.
- [48] J.S. Lim and A.V. Oppenheim, "Enhancement and Band-Width Compression of Noisy Speech," *Proceedings Of The IEEE*, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.
- [49] J.W. Seok and K.S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 1323-1326, 1997.
- [50] J.H.L. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, Vol. 39, pp. 795-805, Apr. 1991.
- [51] J.H.L. Hansen, B. Pellom, "Text-Directed Speech Enhancement Using Phoneme Classification and Feature Map Constrained Vector Quantization" *Speech Communications*, Vol. 21, pp. 169-189, Apr. 1997.
- [52] J.H.L. Hansen, S. Nandkumar, "Robust Estimation of Speech in Noisy Backgrounds Based on Aspects of the Auditory Process," *The Journal of the Acoustical Society of America*, Vol. 97, No. 7, pp. 3833-3849, Jun. 1995.
- [53] J.D. Gibson, B.Koo and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1732-1742, Aug. 1991.
- [54] J. Durbin, "The Fitting of Time Series Models," *Rev. Int. Stat. Inst.*, Vol. 28, pp. 233-244, 1960.
- [55] J.D. Johnston, "Transform Coding of Audio Signal Using Perceptual Noise Criteria," *IEEE J. Select Areas Commun.*, Vol. 6, pp. 314-323, Feb. 1988.

- [56] J. Segura, C. Benitez, A. de la Torre and A. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR," *Proc. ICSLP02*, Denver, Colorado, pp. 225228, Sep. 2002.
- [57] J. Sladek, A. Zschorn, A. and A. Hashemi-Sakhtsari, "Speech to Text Transcription in Support of Pervasive Computing," *Ubiquitous Computing: Trends, Tricks and Traps*, and the Inaugural Asia Pacific Forum on Pervasive Computing, Australia, 2002.
- [58] K. Hermus, I. Dologlou, P. Wambacq and D. Van Compernelle, "Fully Adaptive SVD-Based Noise Removal for Robust Speech Recognition," *Proc. European Conference on Speech Communication and Technology*, volume V, pages 1951-1954, Budapest, Hungary, September 1999.
- [59] K.K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-87, pp. 177-180, 1987.
- [60] K. Yao, E. Visser, O. W. Kwon and T. W. Lee, "A Speech Processing Front-End with Eigenspace Normalization for Robust Speech Recognition in Noisy Automobile Environments," *EUROSPEECH*, pp. 9-12, Geneva, 2003.
- [61] L. Deng, J. Droppo, and A. Acero. "A Bayesian Approach to Speech Feature Enhancement Using the Dynamic Cepstral Prior, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing* ICASSP-02, Vol. I, pp. 829-832, Orlando, Florida, May 2002.
- [62] L.E. Baum, T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164-171, 1970.
- [63] J.C.R. Licklider and I. Pollack, "Effects of Differentiation, Integration and Infinite Peak Clipping Upon the Intelligibility of Speech," *J. Acoust. Soc. Amer.*, Vol. 20, pp. 42-51, 1947.

- [64] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.
- [65] L.R. Rabiner and B.H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, pp. 4-16, Jan. 1986.
- [66] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [67] M. Bahoura and J. Rouat, "Wavelet Speech Enhancement Using the Teager Energy Operator," *IEEE Signal Processing Letters*, Vol. 8, No. 1, pp. 1012, Jan. 2001.
- [68] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-79, pp. 208-211, Apr. 1979.
- [69] M. Deisher and A. Spanias, "Speech Enhancement Using State-bases Estimation and Sinusoidal Modeling," *Journal of Acoustical Society of America*, Vol. 102 (2), pp. 1141-1148, Aug. 1997.
- [70] M. Klein and P. Kabal, "Signal Subspace Speech Enhancement with Perceptual Post-filtering," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-02, Vol. 1, pp. 537-540, May 2002.
- [71] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *J. Acoust. Soc. Am.*, Vol. 66, pp. 1647-1652, Dec. 1979.
- [72] M.H.Hayes, "Statistical Digital Signal Processing and Modeling," Nwe York: John Wiley & Sons, Inc., 1996.
- [73] M.R. Weiss, E. Aschkenasy, and T.W. Parsens, "Study and Development of the INTEL Technique for Improving Speech Intelligibility," *Report NSC-FR/4023*, Nicolet Scientific Corporation, Dec. 1974.

- [74] M.R. Sambur, "Adaptive Noise Cancelling for Speech Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 26, pp. 419-423, Oct. 1978.
- [75] M. Shozakai, S. Nakamura and K. Shikano, "Robust Speech Recognition in Car Environments", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-98, Vol. 5, pp. 269-272, 1998.
- [76] N. J. Miller, Recovery of singing voice from noise by synthesis, Thesis Tech. Rep., ID UTEC-CSC-74-013, May 1973, Univ. Utah, Computer Science Library, Salt Lake City, UT.
- [77] N. Levinson, "The Wiener RMS (Root-Mean-Square) Error Criterion in Filter Design and Prediction," *J.Math Phys.*, Vol. 25, pp. 261-278, 1947.
- [78] N. Ma, M. Bouchard and R. A. Goubran, "A Perceptual Kalman Filtering-Based Approach for Speech Enhancement," *International Symposium for Signal Processing and Applications*, ISSPA, Jul. 2003.
- [79] N. Ma, M. Bouchard and R. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-2004, Vol. 1, pp. 717-720, May 2004.
- [80] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No.2, pp. 126-137, Mar. 1999.
- [81] O. Cappé, "Elimination of The Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 345-349, 1994.
- [82] *Methods for subjective determination of transmission quality*. ITU-T Rec. P.800, Aug. 1996.

- [83] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. ITU-T Rec. P.862, Feb. 2001.
- [84] Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proceedings of the IEEE*, pp. 451-513,, Vol. 88, No. 4, Apr. 2000.
- [85] P. Vary, "Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits," *Signal Processing*, Vol. 8, pp. 387-400, 1985.
- [86] P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.
- [87] Petr Pollak, Pavel Sovka and Jan Uhlik, "Noise Suppression System for a Car", *Proc. of the 3rd European Conference on Speech Communication and Technology*, EUROSPEECH'93, pp.1073-1076, Berlin, Germany, Sep 1993.
- [88] R. J. Ross, R.P.S. ODonoghue and G.M.P. OHare, "Improving Speech Recognition on a Mobile Robot Platform through the use of Top-Down Visual Queues," *18th International Joint Conference on Artificial Intelligence*, IJCAI 03, S116, Mexico, Aug. 2003.
- [89] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp.137-145, Apr. 1980.
- [90] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, pp. 504-512, Jul. 2001.
- [91] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Eur. Signal Processing Conf.*, pp. 1182 - 1185, 1994.

- [92] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 5, pp. 504-512, Jul. 2001.
- [93] R.P. Hellman, "Asymmetry of Masking between Noise and Tone," *Perception and Psychophysics*, Vol. 11, pp.241-246, 1972.
- [94] S.E. Levinson, "Structural methods in Automatic speech Recognition," *Proceedings IEEE*, Vol. 73, No. 11, pp. 1625-1650, Nov. 1985.
- [95] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [96] S. Haykin, "*Adaptive Filter Theory*," 4th edition, Prentice Hall Information and System Sciences Series, 2002.
- [97] S. Haykin and L. Li, "Modified Kalman Filter", *IEEE Trans. on Signal Processing*, Vol 42, No. 5 May 1994
- [98] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-27, No.2, pp. 113-120, Apr. 1979.
- [99] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms," *IEEE Trans. on Speech and Audio Processing*, Vol. 6 No. 4 pp 373-385, Jul 1998.
- [100] S.H. Jensen, P.C. Hansen, S.D. Hansen, and J.A. Srensen, "Reduction of Broad-band Noise in Speech by Truncated QSVD", *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 439-448, Nov.1995.
- [101] V. Stahl, A. Fisher and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-00, 2000.

- [102] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, "Objective Measures of Speech Quality," Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [103] S.S. Stevens, "Calculation of the loudness of complex tones," *Journal of the Acoustical Society of America*, No. 28, pp. 807-832, 1956.
- [104] S. Wang, A. Sekey and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. on Select. Areas in Comm.*, Vol. SAC-10, pp. 819-829, 1992.
- [105] T.Q. Nguyen, "Near-Perfect-Reconstruction Pseudo-QMF Bank," *IEEE Trans. Signal Processing*, vol. 42, pp. 65-76, Jan. 1994.
- [106] T. Ramabadran, J. Meunier, M. Jasiuk and B. Kuser, "Enhancing Distributed Speech Recognition with back-end speech reconstruction," *EUROSPEECH*, pp. 1859-1862, Sep. 2001.
- [107] U. Mittal and N. Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 159-167, Mar. 2000.
- [108] V. Stouten, H. Van hamme, K. Demuynck and P. Wambacq, "Robust Speech Recognition using Model-Based Feature Enhancement," *EUROSPEECH*, pp. 17-20, Geneva, 2003.
- [109] W. T. Hong, "A Discriminative and Robust Training Algorithm for Noisy Speech Recognition," *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 8-11, 2003.
- [110] W.R. Wu and P.C. Chen, "Subband Kalman Filtering for Speech Enhancement," *IEEE Trans. on Circuits and SystemsII: Analog and Digital Signal Processing*, Vol. 45, no. 8, pp. 1072-1083, Aug. 1998.
- [111] W. Yang, M. Benbouchta and R. Yantorno, "Performance of the Modified Bark Spectral Distortion as an Objective Speech Quality Measure," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-98, Vol. 1, pp. 541-544, Seattle, 1998.

- [112] Y. P. Lin and P. P. Vaidyanathan, "Linear Phase Cosine Modulated Maximally Decimated Filter Banks with Perfect Reconstruction," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 2525-2539, Nov. 1995.
- [113] X. Huang et. al., "Mipad, A Multimodel Interaction Prototype," *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 9-12, May 2001.
- [114] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-93, Vol. 2, pp. 355-358, Apr. 1993.
- [115] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.
- [116] Y. Ephraim, D. Malah, and B. H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP, 37, pp.1846-1856, Dec. 1989.
- [117] Y. Ephraim, H. L. Van Trees. "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 4, pp. 251-266, Jul. 1995.
- [118] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-33, No. 2, pp. 443-445, Apr.1985.
- [119] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.
- [120] Y. Hu and P.C. Loizou, "A Subspace Approach for Enhancing Speech Corrupted by Colored Noise," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP-02, Vol. 1, pp. 573-576, May 2002.

- [121] Y. Hu and P.C. Loizou, "A Perceptually Motivated Subspace Approach for Speech Enhancement," *International Conference on Spoken Language Processing*, pp. 1797-1800, 2002.
- [122] Z. Goh, K. C. Tan and B. T. G Tan, "Kalman - Filtering Speech Enhancement Method based on a voiced-unvoiced speech model," *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 5, pp. 510-524, Sep. 1999.