Speech Detection, Enhancement and Compression for Voice Communications

Yong Duk Cho

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

UniS

Centre for Communication Systems Research
School of Electronics, Computing and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

October 2001

© Yong Duk Cho 2001
Summary

Speech signal processing for voice communications can be characterised in terms of silence compression, noise reduction, and speech compression. The limit in the channel bandwidth of voice communication systems requires efficient compression of speech and silence signals while retaining the voice quality. Silence compression by means of both voice activity detection (VAD) and comfort noise generation could present transparent speech-quality while substantially lowering the transmission bit-rate, since pause regions between talk spurts do not include any voice information. Thus, this thesis proposes smoothed likelihood ratio-based VAD, designed on the basis of a behavioural analysis and improvement of a statistical model-based voice activity detector.

Input speech could exhibit noisy signals, which could make the voice communication fatiguing and less intelligible. This task can be alleviated by noise reduction as a preprocessor for speech coding. Noise characteristics in speech enhancement are adapted typically during the pause regions classified by a voice activity detector. However, VAD errors could lead to over- or under-estimation of the noise statistics. Thus, this thesis proposes mixed decision-based noise adaptation based on a integration of soft and hard decision-based methods, defined with the speech presence uncertainty and VAD result, respectively.

At low bit-rate speech coding, the sinusoidal model has been widely applied because of its good nature exploiting the phase redundancy of speech signals. Its performance, however, can be severely smeared by mis-estimation of the pitch. Thus, this thesis proposes a robust pitch estimation technique based on the autocorrelation of spectral amplitudes. Another important parameter in sinusoidal speech coding is the spectral magnitude of the LP-residual signal. It is, however, not easy to directly quantise the magnitudes because the dimensions of the spectral vectors are variable from frame to frame depending on the pitch. To alleviate this problem, this thesis proposes mel-scale-based dimension conversion, which converts the spectral vectors to a fixed dimension based on mel-scale warping. A predictive coding scheme is also employed in order to exploit the inter-frame redundancy between the spectral vectors.

Experimental results show that each proposed technique is suitable for enhancing speech quality for voice communications. Furthermore, an improved speech coder incorporating the proposed techniques is developed. The vocoder gives speech quality comparable to TIA/EIA IS-127 for noisy speech whilst operating at lower than half the bit-rate of the reference coder.

Key words: voice activity detection, speech enhancement, pitch, spectral magnitude quantisation, low bit-rate coding
Acknowledgements

I would like to present my sincere gratitude to my Ph.D. supervisor Prof. Almut Kondoz for offering the opportunity of my Ph.D. study and being the source of professional and personal supports. I also wish to thank to my Ph.D. examiners Prof. W. Balachandran and Dr. Peter Sweeney for their comments and kind advice on my thesis.

Many thanks to all multimedia group members, especially Akin Cellatoglu, Khaldoon Al-Naimi and Stephane Villette. They encouraged me when I was depressed. They were delighted more than me when I was happy. I will remember their sincere friendship and the valuable times shared together.

I would like to thank my parents and parents-in-law, Kyung Mo Choi, Hyung Ki Cho, Jung Ja Jung and Sang Tae Son, whose unconditional love and support helped me make a dream come true. Last but not least, I wish to thank to my wife So Young Son and my son Ha Hyun Cho for being patient and loving me always.
Contents

Contents ......................................................................................................................... i
List of Figures ............................................................................................................ iv
List of Tables ...................................................................................................................viii

1 Introduction .............................................................................................................. 1
  1.1 Background ......................................................................................................... 1
  1.2 Objective and Original Contributions ............................................................ 3
  1.3 Thesis Organisation ......................................................................................... 5

2 A Review of Speech Coding Standards ................................................................ 7
  2.1 Voice Communications over Telecommunication Networks ....................... 7
  2.2 Standard Speech Coders ................................................................................... 9
    2.2.1 ITU-T Speech Coding Standard ........................................................ 9
    2.2.2 European Digital Cellular Telephony ............................................... 11
    2.2.3 North American Digital Cellular Telephony ................................... 11
    2.2.4 Secure Communication Telephony ................................................... 12
    2.2.5 Satellite Telephony ............................................................................... 14
  2.3 Summary ............................................................................................................. 14

3 Voice Activity Detection ....................................................................................... 16
  3.1 Overview of Silence Compression ..................................................................... 16
  3.2 Benefits of Silence Compression ...................................................................... 17
  3.3 Problems in VAD ............................................................................................... 19
  3.4 Standard VAD Methods .................................................................................. 20
    3.4.1 Classification ......................................................................................... 20
    3.4.2 ITU-T G.729B/G.723.1A VAD ........................................................ 21
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.3 ETSI GSM-FR/HR/EFR VAD</td>
<td>23</td>
</tr>
<tr>
<td>3.4.4 ETSI AMR VAD</td>
<td>23</td>
</tr>
<tr>
<td>3.4.5 TIA/EIA IS-127/733 VAD</td>
<td>25</td>
</tr>
<tr>
<td>3.4.6 Experimental Results</td>
<td>27</td>
</tr>
<tr>
<td>3.5 VAD based on a Smoothed Statistical Likelihood Ratio</td>
<td>38</td>
</tr>
<tr>
<td>3.5.1 Motivation</td>
<td>38</td>
</tr>
<tr>
<td>3.5.2 Description of Decision Rules based on Likelihood Ratio Test</td>
<td>38</td>
</tr>
<tr>
<td>3.5.3 Analysis and Improvement of the Likelihood Ratio</td>
<td>39</td>
</tr>
<tr>
<td>3.5.4 Noise Estimation based on the SLR</td>
<td>43</td>
</tr>
<tr>
<td>3.5.5 Experimental Results</td>
<td>44</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>46</td>
</tr>
<tr>
<td>4 Speech Enhancement</td>
<td>48</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>48</td>
</tr>
<tr>
<td>4.2 Review of STSA-based Speech Enhancement</td>
<td>50</td>
</tr>
<tr>
<td>4.2.1 Problem Definition</td>
<td>51</td>
</tr>
<tr>
<td>4.2.2 Spectral Subtraction</td>
<td>52</td>
</tr>
<tr>
<td>4.2.3 Maximum-likelihood Spectral Amplitude Estimation</td>
<td>55</td>
</tr>
<tr>
<td>4.2.4 Wiener Filtering</td>
<td>55</td>
</tr>
<tr>
<td>4.2.5 MMSE Spectral Amplitude Estimation</td>
<td>57</td>
</tr>
<tr>
<td>4.2.6 Spectral Estimation based on the Uncertainty of Speech Presence</td>
<td>58</td>
</tr>
<tr>
<td>4.2.7 Experimental Results</td>
<td>59</td>
</tr>
<tr>
<td>4.2.8 Discussions</td>
<td>61</td>
</tr>
<tr>
<td>4.3 Mixed Decision-based Noise Adaptation</td>
<td>75</td>
</tr>
<tr>
<td>4.3.1 Hard Decision-based Noise Adaptation</td>
<td>75</td>
</tr>
<tr>
<td>4.3.2 Soft Decision-based Noise Adaptation</td>
<td>76</td>
</tr>
<tr>
<td>4.3.3 Mixed Decision-based Noise Adaptation</td>
<td>77</td>
</tr>
<tr>
<td>4.3.4 Experimental Results</td>
<td>77</td>
</tr>
<tr>
<td>4.4 Summary</td>
<td>78</td>
</tr>
</tbody>
</table>
5 Speech Model Parameter Estimation and Quantisation

5.1 A Speech Production Model ................................................................. 85
5.2 Redundancy in Speech ................................................................. 86
5.3 Classical Low Bit-rate Speech Coders .............................................. 87
5.4 Sinusoidal Model-based Speech Coding ............................................. 90
  5.4.1 Speech Modelling ................................................................. 90
  5.4.2 LP-coefficient Estimation ..................................................... 91
  5.4.3 Pitch Estimation ................................................................. 93
  5.4.4 Voicing Estimation ............................................................. 93
  5.4.5 LP-coefficient Quantisation .................................................. 95
  5.4.6 Spectral Magnitudes Quantisation ........................................... 98
5.5 Pitch Estimation using Spectral Autocorrelation ................................. 99
  5.5.1 Introduction ........................................................................... 99
  5.5.2 Spectral Autocorrelation ....................................................... 100
  5.5.3 Pitch Estimation using Spectral Autocorrelation ....................... 103
    5.5.3.1 A PDA based on Spectro-Temporal Autocorrelation .... 103
    5.5.3.2 A PDA based on Spectral Synthesis incorporating Spectral Autocorrelation .................. 108
  5.5.3.3 Spectral Flattening ........................................................... 109
  5.5.4 Experimental Results ............................................................ 110
    5.5.4.1 Analysis of the STA Weighting Factor ......................... 111
    5.5.4.2 Analysis of the SS-SA Weighting Factor ...................... 111
    5.5.4.3 Analysis of the Formant Weighting Factor ................. 114
    5.5.4.4 Performance Evaluation under Background Noise Environ­
          ments ............................................................................... 114
    5.5.4.5 Computational Complexity ........................................... 123
5.6 Predictive and Mel-scale Binary Vector Quantisation ........................... 125
  5.6.1 Introduction ........................................................................... 125
  5.6.2 Distortion Measure ............................................................. 125
  5.6.3 Predictive Quantisation ....................................................... 127
  5.6.4 Mel-Scale Binary Vector Quantisation .................................... 127
  5.6.5 Complexity Optimisation .................................................... 128
5.6.6 Experimental Results .................................................. 129
5.7 Switched Predictive Mel-scale-based Vector Quantisation ........... 131
5.7.1 Introduction .................................................................. 131
5.7.2 Switched-Predictive Mel-scale-based Vector Quantisation .... 133
5.7.3 Experimental Results .................................................... 136
5.8 Summary ........................................................................ 139

6 An Enhanced Split-band LPC Vocoder ................................. 142
6.1 Introduction .................................................................. 142
6.2 Overall Structure of the eSB-LPC Speech Encoder and Decoder .... 143
6.3 Preprocessing ................................................................. 147
6.4 Voice Activity Detection and Noise Reduction ....................... 147
6.5 Estimation of Model Parameters ....................................... 147
6.6 Quantisation of Model Parameters ..................................... 149
6.7 Packetisation and Frame Erasure Handling .......................... 150
6.8 Speech Synthesis ........................................................... 152
6.9 Comfort Noise Generation and Insertion ............................ 154
6.10 Subjective Speech-Quality Evaluation ............................... 154
6.11 Summary ...................................................................... 155

7 Conclusions and Future Research .......................................... 158
7.1 Conclusions ................................................................. 158
7.2 Future Research ........................................................... 160

A List of Abbreviations ......................................................... 162

B List of Publications .......................................................... 166

Bibliography ....................................................................... 168
List of Figures

3.1 An example showing voice active and inactive regions .............................................. 17
3.2 Overall structure of a speech coding system with silence compression. .................. 18
3.3 Block diagram of ITU-T G.729B VAD .............................................................. 22
3.4 Block diagram of ETSI GSM-EFR VAD .......................................................... 23
3.5 Block diagram of ETSI AMR VAD option 1 ..................................................... 24
3.6 Block diagram of ETSI AMR VAD option 2 ..................................................... 25
3.7 Block diagram of TIA/EIA IS-127 RDA ........................................................ 26
3.8 Comparison of speech detection error rates against various vehicle noise levels ....... 28
3.9 Comparison of silence detection error rates against various vehicle noise levels ......... 29
3.10 Comparison of speech detection error rates against various babble noise levels ........ 30
3.11 Comparison of silence detection error rates against various babble noise levels ........ 31
3.12 Comparison by examples of VAD results over vehicle noise of 5 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2 ............................................................ 32
3.13 Comparison by examples of VAD results over vehicle noise of 15 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2 ............................................................ 33
3.14 Comparison by examples of VAD results over vehicle noise of 25 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2 ............................................................ 34
3.15 Comparison by examples of VAD results over babble noise of 5 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2 ............................................................ 35
3.16 Comparison by examples of VAD results over babble noise of 15 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2 ............................................................ 36
List of Figures

3.17 Comparison by examples of VAD results over babble noise of 25 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2. ............................... 37

3.18 Likelihood ratio versus a priori SNR versus a posteriori SNR. The solid lines from top-most to bottom are a posteriori SNRs of 15, 10, 5, 0, -5, -10, -15 dB, respectively. ................................................. 40

3.19 An example of the computed LR (solid line) and SLR (dotted line) of a segment of vehicle noisy signals of 5 dB SNR. The dotted horizontal-line indicates the VAD threshold. The boxed regions in (a) are enlarged in (b) and (c). ........................................................................... 42

3.20 Analysis of the smoothing factor $\kappa$ of the SLR with respect to detection error rates. The noise level is 10 dB SNR, and the noise sources are (a) vehicle and (b) babble. .................................................. 45

4.1 Block diagram of general STSA-based speech enhancement methods. ... 53

4.2 Comparison of SEGSNR improvements of STSA-based speech enhancement methods over vehicle noise environments. .................................................. 62

4.3 Comparison of ISDs of STSA-based speech enhancement methods over vehicle noise environments. .................................................. 63

4.4 Comparison of SEGSNR improvements of STSA-based speech enhancement methods over helicopter noise environments. .................................................. 64

4.5 Comparison of ISDs of STSA-based speech enhancement methods over helicopter noise environments. .................................................. 65

4.6 Noisy speech (a) waveform and (b) spectrogram. The noise source and level are vehicle and 5 dB SNR, respectively. .................................................. 66

4.7 Noise-free speech (a) Waveform and (b) spectrogram. .................................................. 67

4.8 Enhanced speech (a) Waveform and (b) spectrogram by theoretical limit. .................. 68

4.9 Enhanced speech (a) Waveform and (b) spectrogram by PSS. .................................................. 69

4.10 Enhanced speech (a) Waveform and (b) spectrogram by GBSS with $\nu = 2$, $\alpha = 4$, and $\beta = 0.1$. .................................................. 70

4.11 Enhanced speech (a) Waveform and (b) spectrogram by ML-STSA estimation. .................................................. 71

4.12 Enhanced speech (a) Waveform and (b) spectrogram by WF. .................................................. 72

4.13 Enhanced speech (a) Waveform and (b) spectrogram by MMSE-STSA estimation. .................................................. 73

4.14 Enhanced speech (a) Waveform and (b) spectrogram by MMSE-STSA estimation with speech presence uncertainty. .................................................. 74

4.15 Performance comparison in terms of SEGSNR improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 0 dB SNR. 79
4.16 Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 5 dB SNR. 80
4.17 Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 10 dB SNR. 81
4.18 Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 0 dB SNR. 82
4.19 Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 5 dB SNR. 83
4.20 Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 10 dB SNR. 84

5.1 An example of (a) speech signal of $T_0 = 34$-sample ($F_s = 8$ kHz), (b) magnitude spectrum, (c) zero-crossing spectrum, and (d) spectral autocorrelation ............................................................................................ 102
5.2 An example of (a) speech signal of $T_0 = 59$-sample ($F_s = 8$ kHz), (b) magnitude spectrum, (c) zero-crossing spectrum, and (d) spectral autocorrelation ................................................................................................................. 104
5.3 Comparison of TA, SA, and STA ($\alpha = 0.5$) for (a) high, (b) mid, and (c) low-pitched speech signals with pitch periods 32, 59, and 100 samples, respectively. ............................................................................................................ 107
5.4 Comparison between SS and SS-SA ($\beta = 0.25$) for high, mid, and low-pitched signals shown in Fig. 5.3(a), (b), and (c), respectively. The SAs are shown in Fig. 5.3(g), (h), and (i) ................................................................................. 109
5.5 Analysis of the effect of the STA weighting factor $\alpha$ in terms of the pitch error rate. Here, the formant weighting factor $\gamma$ is 0.9. .................................... 112
5.6 Analysis of the effect of the SS-SA weighting factor $\beta$ in terms of the pitch error rate. Here, the formant weighting factor $\gamma$ is 0.9 ......................... 113
5.7 Analysis of the effect of the formant weighting factor $\gamma$ in terms of the pitch error rate. Here, $\alpha$ and $\beta$ defined in the STA and SS-SA functions are 0.5 and 0.25, respectively; and the rectangular window is applied to TA calculation in the STA .................................................................................... 114
5.8 Analysis of the effect of the weighting factor $\theta$ in WTA. .............................. 115
5.9 Comparison of pitch contours of PDAs for (a) noise-free female speech. (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs. .......................................................... 119
5.10 Comparison of pitch contours of PDAs for (a) noise-free male speech (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs. .......................................................... 120
5.11 Comparison of pitch contours of PDAs for (a) female speech corrupted by vehicle noise of 15 SNR dB. (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs. 121
5.12 Comparison of pitch contours of PDAs for (a) male speech corrupted by vehicle noise of 15 SNR dB. (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs ................................................122

5.13 The relationship between the spectral lag $\omega_r$ (the number of frequency bins) and the temporal lag $\tau$ (the number of time-domain samples) for the DFT size $M = 256$. .......................................................... 124

5.14 Block diagram of predictive and mel-scale binary vector quantiser. ...... 126

5.15 Equivalent WSNR (= 13.0 dB) relation against the dimension of PM-BVQ and the order of LP-analysis. ............................................................ 131

5.16 Performance comparison between PMBVQ and a high order LP-model against the pitch. The order of the high order LP-model is 26. The vector dimension and LP-order for PMBVQ are 12 and 10, respectively. 132

5.17 Distribution of the pitch of the speech material used in the test. .... 132

5.18 Analysis of the inter-frame correlation of LP-residual magnitude-spectra: (a) waveform of a female speech segment, (b) developed, every 10-ms, LP-residual magnitude-spectra of the boxed speech in (a), (c) normalised correlation ratio of the frame-to-frame harmonic sequence in (b), and (d) overlapped display of (b). .............................................................. 134

5.19 Block diagram of SP-MVQ encoder. .................................................... 135

5.20 Performance comparison between ML-MVQ, P-MVQ, and SP-MVQ of which the numbers of allocated bits are 14, 14, and 13+1(VQ switch-mode), respectively. .......................................................... 137

6.1 Block diagram of the eSB-LPC codec. .................................................. 144

6.2 Block diagram of the eSB-LPC encoder. .............................................. 145

6.3 Block diagram of the forward and inverse packetisation of the eSB-LPC codec. ................................................................. 146
List of Tables

2.1 ITU-T narrow-band voice coding standards ................................................ 10
2.2 ETSI voice coding standards for GSM mobile communications ............ 12
2.3 TIA/EIA voice coding standards for North American CDMA/TDMA mobile communications ................................................................................... 13
2.4 DoD voice coding standards ............................................................................ 13
2.5 INMARSAT voice coding standards ............................................................ 14

3.1 Classification of standard VAD methods depending on input features. The values in parentheses indicate the number of spectral sub-bands.  .......... 21
3.2 Comparison of speech and silence detection error rates among SLR-based, LR-based, AMR2, and G.729B VADs. The LR+HO means LR-based VAD with the hangover scheme. ................................. 47

5.1 Classification of speech coders according to redundancy features being exploited. .......................................................... 90
5.2 Comparison of the number of strong peaks of periodic impulse-train-like signals. ................................................................. 106
5.3 Comparison of pitch error rates among the STA, WTA, TA, SS-SA, WSS, and SS-based PDAs under vehicle noise environments. ............... 117
5.4 Comparison of pitch error rates among the STA, WTA, TA, SS-SA, WSS, and SS-based PDAs under babble noise environments ............... 118
5.5 Comparison of the computational complexity. ........................................... 123
5.6 Comparison of WSNR(dB) between the PMBVQ and the LP-only methods. .................................................................................. 130
5.7 Bits allocation of the SB-LPC vocoder with SP-MVQ. The frame size is 20 ms, and the frame is split into two subframes. ......................... 138
5.8 Preference test results. ................................................................................ 140

6.1 Bits allocation of the eSB-LPC vocoder for voice-active frames. ............ 151
List of Tables

6.2 Bits allocation of the eSB-LPC vocoder for CNI frames. .... 151
6.3 Packet structure of the eSB-LPC vocoder. ................. 151
6.4 Mean opinion score. ........................................... 155
6.5 Subjective speech-quality test results in term of the MOS. . 156
6.6 Comparison of the average bit-rates between the EVRC and eSB-LPC speech coders with the VAD mode on. ................. 156

x
Chapter 1

Introduction

1.1 Background

Speech is one of the most natural and fundamental forms of communications among human beings. The advance in telecommunication technologies realises convenient manners for voice communications through wireless and wired lines. Furthermore, the digital technology presents relatively high and consistent speech-quality over telecommunication networks. In order to provide that performance over a large population of people, significant reduction in the channel bandwidth is of great concern in most speech communication systems. In mobile communications, the task is vital not only to increase spectral efficiency but also to reduce wireless channel interference. Speech coding thus plays an important role in high quality communications. Moreover, recent interests in voice communications over packet networks increase the need for efficient speech coding in order to mitigate quality degradation due to network congestion.

Speech coding including speech compression, silence compression, and noise reduction is a crucial technique in bandwidth-limited communication systems. Hence, it is a meaningful work, developing advanced speech signal processing algorithms for high quality voice communications.

Voice activity detection (VAD) is a classification technology between speech and silence regions. A signal can be classified into speech, if it includes a speech component, or
1.1. Background

Silence compression by means of VAD can substantially increase the channel capacity since the talk spurt in typical two-way voice communications occupies just about 30 ~ 40%. Thus, VAD techniques are widely applied to digital cellular telephone systems in the forms of variable bit-rate (VBR) speech coding in code-division multiple access (CDMA) systems, and discontinuous transmission (DTX) in the global system for mobile communications (GSM). The interest in voice communications over packet networks also increases the need for VAD due to its cost-effective nature, producing transmission packets with payload just during talk spurs.

Input signals for voice communications typically exhibit noisy speech corrupted by environmental noise sources. The noise would make the communication fatiguing and degrade the voice intelligibility. Thus, noise reduction suppressing the noise components in the input signal can present improved voice communication quality. The task of speech enhancement has two kinds of objectives: improvement of perceptual speech-quality against background noise, and improvement of the accuracy of model-parameter estimation for speech coders.

Speech compression is a major part in coding the speech signal. Depending on the bit-rate, it is typically classified into high, medium, low, and very low bit-rates, each of which ranges 64 ~ 16, 16 ~ 8, 8 ~ 1, and below 1 kbit/s, respectively. Many speech coders have been developed and standardised over the various bit rates for the last two decades [1] [2] [3]. High bit rate coders are mainly applied to fixed networks such as digital circuit multiplexing equipments. Most of modern medium bit-rate coders are mainly developed for the last decade with the boom of the cellular telephony. Recently many research works focus on developing low bit-rate coders, especially for lower than 4 kbit/s, so as to further reduce the bit-rate while maintaining high quality speech.

Technical approaches for low bit-rate speech coding can be classified into waveform and parametric coding methods. Waveform coders such as code-excited linear prediction (CELP) coders try to faithfully represent the original waveform shape, but parametric coders do not. Waveform coders are widely applied for bit rates higher than 4 kbit/s. On the other hand, for bit rates lower than 4 kbit/s, parametric coders such as sinusoidal speech coders are typically selected. Parametric coders represent the speech signal with a small number of model parameters based on a speech production model.
1.2 Objective and Original Contributions

This thesis aims to achieve enhanced voice communication quality through advanced speech signal processing techniques composed of improved VAD, speech enhancement, and speech compression methods. The goal is accomplished by the following original contributions:

- **Improved VAD based on a smoothed statistical likelihood ratio**: This research presents the behavioural mechanism of a statistical model-based VAD method, featuring a likelihood ratio test for the activity decision. From investigation of the VAD function, it is found that detection errors could occur frequently at speech offset regions because of the delay term in the decision-directed parameter estimator, employed for the estimation of an unknown parameter of the likelihood ratio. Hence, this thesis proposes a smoothed likelihood ratio so as to alleviate the detection errors at the offset region. Objective test results show that the proposed scheme is useful for achieving a considerable performance improvement for VAD. Additionally, the proposed VAD method gives detection performances superior to G.729B VAD and comparable to AMR VAD option 2.

- **Speech enhancement based on mixed decision-based noise adaptation**: In speech enhancement, the noise statistics are conventionally adapted by a hard decision method in cooperation with VAD. As a result, the quality of enhanced speech can be affected by the performance of VAD. A mixed decision method is proposed for noise adaptation in speech enhancement so as to give robustness against VAD errors by conducting the adaptation with the combination of soft and hard decision-based techniques. Objective speech-quality tests, in terms of the segmental SNR improvement and the Itakura-Saito distortion, demonstrate its superiority in comparison with both hard and soft decision-based methods.

- **Robust pitch estimation using spectral autocorrelation**: Incorrect detection of the pitch multiples for voiced speech is a well-known problem in time-domain autocorrelation (TA)-based pitch determination algorithms (PDA). However, pitch estimation based on spectral autocorrelation (SA) does not exhibit
maximal peaks for the pitch multiples but can give rise to submultiples. Hence, this thesis proposes a PDA based on spectro-temporal autocorrelation (STA), defined a weighted sum between SA and TA for each lag. The STA function presents robustness for pitch estimation due to the property that each term, SA and TA, in STA can compensate for pitch multiples and submultiples. STA is capable of emphasising the feature of the pitch while those of its multiples and submultiples are suppressed. It is also possible to alleviate the pitch-multiple problem in spectral synthesis (SS)-based PDAs, widely used in sinusoidal speech coders, with the aid of SA. Pitch detection experiments show that the STA-based PDA results in performances, in terms of the gross pitch-error rate, superior to conventional ones. The SS-based PDA combined with SA also gives performances much better than conventional ones, but slightly worse than the STA-based method.

- Efficient quantisation of variable dimension spectral magnitudes using predictive and mel-scale-based coding schemes: This thesis proposes two kinds of spectral magnitude quantisation techniques for sinusoidal speech coders. One is predictive mel-scale binary vector quantisation (PMBVQ), and the other is switched-predictive mel-scale vector quantisation (SP-MVQ). Both of them exploit the intra-frame perceptual preference of the magnitudes using a mel-scale-based warping technique, and the inter-frame statistical correlation of the successive magnitude vectors using a predictive coding scheme. Subjective speech-quality tests with SP-MVQ show its suitability in developing high quality sinusoidal speech coders.

- Design of an enhanced split-band LPC vocoder: A new speech coder is designed on the baseline structure of a split-band linear predictive coding (SB-LPC) vocoder. The proposed coder, named an enhanced split-band LPC (eSB-LPC) vocoder, features new VAD, speech enhancement, pitch estimation, and spectral magnitude quantisation of which the algorithmic characteristics are described above. The eSB-LPC vocoder operates at the maximum bit-rate of 4 kbit/s, but at the average rate of about 2.7 kbit/s. It produces sound quality comparable to the TIA/EIA IS-127 coder for noisy speech. The TIA/EIA IS-127 speech coder
operates at the maximum and average bit rates of 8.55 and about 6.9 kbit/s, respectively.

1.3 Thesis Organisation

The dissertation is organised with the following chapters.

Chapter 2 reviews standard speech coders for voice communications.

Chapter 3 reviews and proposes VAD techniques. It presents the general overview, the advantage of silence compression in voice communication, problems, and comparative studies of the standard VAD methods for ITU-T G.729B, ETSI GSM-EFR, ETSI AMR, and TIA/EIA IS-127. Moreover, this chapter proposes an improved VAD technique based on a smoothed statistical likelihood ratio.

Chapter 4 reviews and proposes speech enhancement techniques against background noise. The review includes comparative studies of noise reduction methods based on enhanced short-time spectral amplitudes estimated by spectral subtraction, maximum likelihood, Wiener filtering, minimum mean square error, or speech presence uncertainty. This chapter proposes a mixed-decision based noise adaptation technique to robustly keep track of noise statistics.

Chapter 5 reviews speech compression techniques, and proposes new methods for enhanced model parameter estimation and quantisation. The review discusses a speech production model, redundancy features in speech signals, classical low bit-rate speech coders, and algorithmic details of sinusoidal speech coders. This chapter proposes three novel techniques for robust pitch estimation and efficient spectral magnitude quantisation.

Chapter 6 presents implementation details of an enhanced sinusoidal speech coder incorporating the proposed techniques in Chapter 3, 4, and 5. The algorithmic description includes the overall encoding and decoding structure, signal preprocessing, VAD, noise reduction, model parameters estimation and quantisation, packetisation with frame erasure handling, and speech synthesis with comfort noise insertion and generation.
1.3. *Thesis Organisation*

Subjective speech quantity test results in terms of the mean opinion score are also presented.

Chapter 7 presents the concluding remarks together with future works.
Chapter 2

A Review of Speech Coding Standards

This chapter reviews standard speech coders for voice communications. Section 2.1 describes telecommunication networks which deliver digital voice. Next, in Section 2.2, speech coders recommended by standard organisations are reviewed. Finally, the summary of this chapter is presented in Section 2.3.

2.1 Voice Communications over Telecommunication Networks

Speech communication systems deliver voice information to remote sites through wired or wireless communication networks. Thus, the history of the voice signal processing techniques has been in line with that of telecommunication systems. In other words, the evolution or introduction of a novel digital telecommunication system requires suitable speech signal processing techniques to meet the service requirements.

Public switched telephone networks (PSTNs) started with analog transmission, and by now it has been made almost completely digital transmission except subscriber connection. The digital PSTN conveys the voice signal in the form of A/µ-Law pulse code modulation (PCM) [4] at the transmission rate of 64 kbit/s with a sampling frequency
of 8 kHz. The capacity of the PSTN can be improved by speech signal compression followed by a multiplexing scheme. For example, in the European standard, an E1 trunk is capable of serving a maximum of 32 subscribers on the carrier frequency 2.048 MHz with A-Law PCM of 64 kbit/s. Speech compression at the rate of 8 kbit/s combined with digital circuit multiplexing enables it to serve 256 subscribers per E1 trunk. Of course, at the other site of the network, a system with both proper demultiplexing and speech decoding should be deployed.

PSTNs transmit the speech in the digital form between base stations only. Thus, speech quality can be distorted because of the analog channel between the subscriber and base station. Integrated services digital networks (ISDNs) present much better speech quality by end-to-end digital transmission together with wide-band speech of 7 kHz signal bandwidth.

Mobile communications based on the digital cellular technology started to appear on the market in the late 1980s. In mobile communications, speech coding techniques play an important role due to the limit in radio channel bandwidth. The coding rate is typically determined to around 13 or 8 kbit/s by European and North American standard organizations. Later, in the mid 1990s, half rate speech coders were introduced to mitigate problems in the regions with high interference. However, the speech quality of the mobile phones could not meet the toll quality until the mid 1990s. Advances in voice coding technologies finally enabled toll quality speech in the late 1990s.

Secure communications, typically used in military systems in emergency, are another important area in wireless voice communications. Intelligibility of the voice is the main concern in secure communications.

The interest in voice transmission over packet networks, typically called voice over internet protocol (VoIP) [5], is increasing due to its cost effective nature in voice communications. The cost for telecommunications can be reduced when speech is transmitted in packet forms because the silence periods do not have to be transmitted. At the moment, most of voice coding techniques for VoIP reuses traditional voice coders. However, the traditional voice coding techniques are developed mostly targeting circuit-based telecommunication systems. It means that there is room for performance improvement
of the packet voice by means of a new vocoder. The packet network typically exhibits the problems of packet loss, delay, and jitter. Thus, a vocoder with the capability of robustness against the packet delivery problems is highly desirable.

Voice over wireless packet networks is a prospective area considering the evolution of mobile communications. Future telecommunication systems suggest a move toward wireless packet networks, due to the boom of the internet and mobile communications. The European telecommunication standard body has standardised packet data transmission over global system for mobile communications (GSM), called general packet radio service (GPRS) [6], and moreover higher bit-rate packet service over GSM, called enhanced data rates for GSM evolution (EDGE) [7]. The third generation mobile communication systems for universal mobile telecommunications service (UMTS) [8] is also based on wireless packet services.

2.2 Standard Speech Coders

Delivery of the voice to a remote site requires efficient speech coding techniques because telecommunication environments are restricted by the channel resources. Thus, efficient speech coding techniques minimising channel overload while serving high voice quality is a great concern in designing voice communication systems. This section reviews standard speech coders mostly developed for specific communication systems.

2.2.1 ITU-T Speech Coding Standard

Traditionally the International Telecommunication Union - Telecommunication (ITU-T, formerly CCITT) has standardised speech coding methods mainly for the PSTN telephony with 4 kHz Nyquist and 8 kHz sampling frequencies, aiming to improve telecommunication network capacity by means of digital circuit multiplexing. Additionally, the ITU-T has been conducted standardisation for wide-band speech coders to support 7 kHz Nyquist and 16 kHz sampling frequencies mainly for the ISDN.

In 1972 the ITU-T released G.711 [4], A/µ-Law PCMs for the 64 kbit/s speech coding standard, which is designed on the basis of logarithmic scaling of each pulse amplitude.
2.2. Standard Speech Coders

<table>
<thead>
<tr>
<th>Speech coder</th>
<th>Bit-rate (kbit/s)</th>
<th>VAD</th>
<th>NR</th>
<th>Delay (ms)</th>
<th>Quality</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>G.711 (A/µ-Law PCM)</td>
<td>64</td>
<td>No</td>
<td>No</td>
<td>0</td>
<td>Toll</td>
<td>1972</td>
</tr>
<tr>
<td>G.726 (ADPCM)</td>
<td>40/32/24/16</td>
<td>No</td>
<td>No</td>
<td>0.25</td>
<td>Toll</td>
<td>1990</td>
</tr>
<tr>
<td>G.728 (LD-CELP)</td>
<td>16</td>
<td>No</td>
<td>No</td>
<td>1.25</td>
<td>Toll</td>
<td>1992</td>
</tr>
<tr>
<td>G.729 (CSA-CELP)</td>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>25</td>
<td>Toll</td>
<td>1996</td>
</tr>
<tr>
<td>G.723.1 (MP-MLQ/ACELP)</td>
<td>6.3/5.3</td>
<td>Yes</td>
<td>No</td>
<td>67.5</td>
<td>Toll</td>
<td>1995</td>
</tr>
<tr>
<td>G.4k (to be determined)</td>
<td>4</td>
<td>-</td>
<td>Yes</td>
<td>~55</td>
<td>Near-toll</td>
<td>2001</td>
</tr>
</tbody>
</table>

G.711 was deployed in PSTNs throughout the world. Since then, the ITU-T has been standardising non-PCM speech coders, named the G.72x series. ITU-T released G.721, the 32 kbit/s adaptive differential pulse code modulation (ADPCM) coder; followed by the extended versions, G.723 and G.726 [9], supporting 24/40 kbit/s and 40/32/24/16 kbit/s, respectively. The latest ADPCM version, G.726, supersedes the former ones, G.721 and G.723. ITU-T speech coders except G.723.1 [10] had been developed, decreasing the bit-rate to a half level. G.728 [11] and G.729 [12] speech coders finalised in 1992 and 1996 were recommended at the rates of 16 and 8 kbit/s, respectively. Additionally, ITU-T released G.723.1 [10], the 5.3/6.3 kbit/s dual-rate speech coder, for video telephony systems. G.728, G.729, and G.723.1 are designed based on code excited linear prediction (CELP) technologies. For discontinuous transmission (DTX), ITU-T released the extended versions of G.729 and G.723.1, called G.729B [13] and G.723.1A [14], respectively, which are widely used in packet-based voice communications [6] due to the silence compression schemes. Now, for 4 kbit/s speech coding, two kinds of coders are proposed to ITU-T. One is based on the CELP model, and the other is a hybrid model of CELP and sinusoidal speech coding technologies [15] [16]. Narrow band standard speech coders recommended by ITU-T are summarised in Table 2.1.

The ITU-T released two kinds of wide-band speech coders, G.722 [17] and G.722.1 [18],
targeting mainly multimedia communications with a high voice quality. G.722 [17] supports three bit-rates, 64, 56, and 48 kbit/s based on sub-band ADPCM (SB-ADPCM). It decomposes input signals into low and high sub-bands using the quadrature mirror filter, and then quantises the band-pass filtered signals using ADPCM with variable step sizes depending on the sub-band. G.722.1 [18] operates at the rates of 32 and 24 kbit/s based on the transform coding technique. Currently, a new wide-band speech coder operating at 13/16/20/24 kbit/s is under standardisation.

### 2.2.2 European Digital Cellular Telephony

Digital cellular telephony triggered active development of a number of speech coders. The European Telecommunications Standards Institute (ETSI) released the GSM full rate (FR) speech coder operating at 13 kbit/s [19]. Since then, the ETSI continued to recommend 5.6 kbit/s GSM half rate (HR) and 12.2 kbit/s GSM enhanced full rate (EFR) speech coders [20] [21]. Recently, the ETSI standardised a new speech coder, called the adaptive multi-rate (AMR) coder [22], operating at eight bit-rates from 12.2 to 4.75 kbit/s. The AMR coder aims to provide enhanced speech-quality based on optimal selection between the source and channel coding schemes. Under high radio interference, AMR is capable of allocating more bits for channel coding rather than source coding; and vice versa.

Additionally, each ETSI speech coder is capable of silence compression [23] [24] [25] [26], which presents not only channel interference reduction but also battery lifetime extension for mobile communications. Standard speech coders for European mobile communications are summarised in Table 2.2.

### 2.2.3 North American Digital Cellular Telephony

In North America, the Telecommunication Industries Association (TIA) / Electronic Industries Association (EIA) conducts standardisation for the mobile communication based on CDMA and time division multiple access (TDMA) technologies. In CDMA systems, silence compression schemes can be employed naturally due to the soft channel
2.2. Standard Speech Coders

Table 2.2: ETSI voice coding standards for GSM mobile communications

<table>
<thead>
<tr>
<th>Speech coder</th>
<th>Bit-rate (kbit/s)</th>
<th>Delay (ms)</th>
<th>VAD</th>
<th>NR</th>
<th>Quality</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR (RPE-LTP)</td>
<td>12</td>
<td>40</td>
<td>Yes</td>
<td>No</td>
<td>Near-toll</td>
<td>1987</td>
</tr>
<tr>
<td>HR (VSELP)</td>
<td>5.6</td>
<td>45</td>
<td>Yes</td>
<td>No</td>
<td>Near-toll</td>
<td>1994</td>
</tr>
<tr>
<td>EFR (ACELP)</td>
<td>12.2</td>
<td>40</td>
<td>Yes</td>
<td>No</td>
<td>Toll</td>
<td>1998</td>
</tr>
<tr>
<td>AMR (ACELP)</td>
<td>12.2/10.2/7.95/7.4/6.7/5.9/5.15/4.75</td>
<td>40/45</td>
<td>Yes</td>
<td>No</td>
<td>Toll</td>
<td>1999</td>
</tr>
</tbody>
</table>

capacity. The TIA/EIA adopted Qualcomm CELP (QCELP) [27] for Interim Standard-96-A (IS-96-A), operating at variable bit rates between 8 and 0.8 kbit/s controlled by a rate determination algorithm. Subsequently, the TIA/EIA released IS-127 [28], the enhanced variable rate coder, featuring a novel function for noise reduction as a preprocessor to the speech compression module. Under noisy background condition, noise reduction provides a more comfortable speech-quality by enhancing noisy speech signals. For personal communication systems, the TIA/EIA released IS-733 [29], which operates at variable bit rates between 14.4 and 1.8 kbit/s. On the other hand, for North American TDMA standards, the TIA/EIA released IS-54 and IS-641-A for FR and EFR speech coding, respectively [30] [31]. Standard speech coders for North American mobile communications are summarised in Table 2.3.

2.2.4 Secure Communication Telephony

Speech coding is a crucial part in secure communications. In this environment, voice intelligibility is a major concern in order to deliver exact voice commands in emergency. Main standardisation has been organised by the department of defense (DoD) in the United States of America. The DoD released Federal Standard-1015 (FS-1015) and FS-
Table 2.3: TIA/EIA voice coding standards for North American CDMA/TDMA mobile communications

<table>
<thead>
<tr>
<th>Speech coder</th>
<th>Bit-rate (kbit/s)</th>
<th>VAD</th>
<th>NR</th>
<th>Delay (ms)</th>
<th>Quality</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS-96-A (QCELP)</td>
<td>8.5/4/2/0.8</td>
<td>Yes</td>
<td>No</td>
<td>45</td>
<td>Near-toll</td>
<td>1993</td>
</tr>
<tr>
<td>IS-127 (EVRC)</td>
<td>8.5/4/2/0.8</td>
<td>Yes</td>
<td>Yes</td>
<td>45</td>
<td>Toll</td>
<td>1995</td>
</tr>
<tr>
<td>IS-733 (QCELP)</td>
<td>14.4/7.2/3.6/1.8</td>
<td>Yes</td>
<td>No</td>
<td>45</td>
<td>Toll</td>
<td>1998</td>
</tr>
<tr>
<td>IS-54 (VSELP)</td>
<td>7.95</td>
<td>Yes</td>
<td>No</td>
<td>45</td>
<td>Near-toll</td>
<td>1989</td>
</tr>
<tr>
<td>IS-641-A (ACELP)</td>
<td>7.4</td>
<td>Yes</td>
<td>No</td>
<td>45</td>
<td>Toll</td>
<td>1996</td>
</tr>
</tbody>
</table>

Table 2.4: DoD voice coding standards

<table>
<thead>
<tr>
<th>Speech coder</th>
<th>Bit-rate (kbit/s)</th>
<th>VAD</th>
<th>NR</th>
<th>Delay (ms)</th>
<th>Quality</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS-1015 (LPC-10e)</td>
<td>2.4</td>
<td>No</td>
<td>No</td>
<td>115</td>
<td>Intelligible</td>
<td>1984</td>
</tr>
<tr>
<td>FS-1016 (CELP)</td>
<td>4.8</td>
<td>No</td>
<td>No</td>
<td>67.5</td>
<td>Communication</td>
<td>1991</td>
</tr>
<tr>
<td>New DoD 2.4 (MELP)</td>
<td>2.4</td>
<td>No</td>
<td>No</td>
<td>67.5</td>
<td>Communication</td>
<td>1996</td>
</tr>
</tbody>
</table>

1016, called 2.4 kbit/s LPC-10e and 4.8 kbit/s CELP coders, respectively [32] [33] [34]. The DoD also standardised a new 2.4 kbit/s speech coder [35], based on the mixed excitation linear prediction (MELP) vocoder [36] using the sinusoidal speech model. The new 2.4 kbit/s DoD speech coder gives speech quality better that 4.8 kbit/s FS-1016. Note that parametric coders, such as MELP, are widely used in secure communications due to its intelligible speech-quality even under 2 kbit/s, while waveform coders typically result in severe quality degradation at low bit rates below 4 kbit/s. The DoD standard speech coders are summarised in Table 2.4.
2.3. Summary

In this chapter, voice communication systems have been reviewed together with standard speech coders, and it has been found that the speech coders are developed mostly based on the characteristics of the applied communication channels. The ITU-T standardised speech coders with the requirement of toll quality speech. The regional standard bodies for the digital mobile telephony have firstly selected near-toll quality speech coders, such as the GSM-FR, TIA/EIA IS-96, and TIA/EIA IS-54. Later, each standardisation body adopted the enhanced vocoder versions, such as the ETSI GSM-EFR, TIA/EIA IS-127, and TIA/EIA IS-641, which present toll quality speech over mobile communication environments. The mobile telephone standards are highly concerned

### Table 2.5: INMARSAT voice coding standards

<table>
<thead>
<tr>
<th>Speech coder</th>
<th>Bit-rate (kbit/s)</th>
<th>VAD</th>
<th>NR</th>
<th>Delay (ms)</th>
<th>Quality</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMBE</td>
<td>4.15</td>
<td>No</td>
<td>No</td>
<td>120</td>
<td>Communication</td>
<td>1990</td>
</tr>
<tr>
<td>AMBE</td>
<td>3.6</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

2.2.5 Satellite Telephony

The international maritime satellite corporation (INMARSAT) has adopted a few speech coders for satellite communications. Because of the distance between the satellite and the mobile terminals, the channel typically exhibits high distortion. Thus the satellite communication requires low bit-rate coders in order to allocate a relatively large number of bits for channel coding. The INMARSAT selected 4.15 kbit/s improved multiband excitation (IMBE) [37] and 3.6 kbit/s advanced multiband excitation (AMBE) vocoders for INMARSAT M and INMARSAT Mini-M systems, respectively. The INMARSAT standard speech coders for satellite communications are summarised in Table 2.5.
with the channel noise effect because the wireless channel for multiple access systems can be easily smeared by fading and interference. Thus, the vocoders are also designed considering the channel effect. In the case of the ETSI GSM, the coders are designed in cooperation of channel coders. On the other hand, TIA/EIA CDMA speech coders employ variable rate schemes so as to minimise channel interference and finally increase the number of users. In secure and satellite communications, the DoD and IMMARSAT selected parametric vocoders, such as LPC-10e, the new 2.4 kbit/s DoD vocoder, IMBE, and AMBE.

At the point of silence compression, the regional standard bodies for mobile communication systems introduced VAD tailored to their multiple access systems. Especially in CDMA-based systems, owing to the soft channel capacity, VAD can be naturally integrated into the system to increase the channel capacity. Considering that future mobile communication systems are based on wideband CDMA, technical advances in VAD will contribute to enhancing mobile system performance.

Noise reduction started to be employed in the TIA/EIA IS-127, and contributed to improve the speech quality under noisy background condition. The recent coder for ITU-T G.4k is expected to adopt speech enhancement as a preprocessor.
Chapter 3

Voice Activity Detection

The objective of this chapter is to achieve a VAD technique with high performances in terms of maximising the detection rate for speech regions while minimising the false alarm rate for silence regions. Section 3.1 defines the task of VAD, and introduces the overall structure of silence compression systems. Section 3.2 discusses the potential benefits gained by silence compression. In Section 3.3, problems in VAD are given. Detailed algorithmic characteristics of standard VAD methods are presented in Section 3.4. A novel VAD method is proposed in Section 3.5. Finally, the summary of this chapter is presented in Section 3.6.

3.1 Overview of Silence Compression

In voice communications, speech can be characterised as a discontinuous media because of the pauses which is a unique feature compared to other multimedia signals, such as video and audio. The regions where voice information exists are called voice-active; and the pauses between talk spurts are called voice-inactive or silence. An example illustrating voice active and inactive regions for a segment of speech signals is shown in Figure 3.1.

VAD is a technique that detects the presence of speech for a segment of input signals. At the decoder, signals for voice-inactive frames can be constructed by means of
3.2. Benefits of Silence Compression

Speech communication systems exploiting silence compression are capable of providing various benefits for bandwidth-limited communication channels. The advantages can be discussed in several points of view depending on application areas, as

- Cochannel interference reduction in cellular communications: It is possible to
3.2. Benefits of Silence Compression

Figure 3.2: Overall structure of a speech coding system with silence compression.

- Suppress cochannel interferences in cell-based wireless communication systems by decreasing transmission energy during pauses.

- Improvement of the soft channel capacity in the code division multiple access (CDMA) system: CDMA systems can be characterised in terms of power control. The CDMA channel capacity is flexible, but the transmission performance can be limited depending on the strength of interference. In other words, the feasible capacity of the CDMA system is more related with the amount of channel interference rather than the number of users. By varying the transmission power, depending on the importance of the signal being transmitted, the cell capacity can be improved. Allocating very low power during pauses compared with talk-spurts, the number of users can be increased in proportion to an increase of pause periods.

- Power saving for mobile terminals: Mobile terminals do not have to transmit radio signals during pauses. Thus, the battery life time of the terminals can be extended by saving the power during silence periods.
• Increase in channel capacity by statistical multiplexing: A channel can be granted just during talk spurts, and be released during pauses. A user occupies a channel, once granted, to the end of a talk spurt. To get the grant again, the user requests an empty channel at the start of the next talk spurt. Thus, the channel resources can be utilised efficiently by the statistical multiplexing scheme, which allows more number of users to communicate at the same time over limited channel resources.

• Reduction in packet losses for packet-based networks: Packet networks are also capable of providing the soft system capacity like CDMA networks. In other words, the channel capacity is limited by the number of packets in the network rather than by that of users. Thus, in packet-based voiced communication systems, the number of users can be increased by silence compression.

• Bit-rate reduction: Substantial reduction in the transmission bit-rate can be attained from speech compression techniques. The silence compression scheme presents additional reduction in the bit-rate regardless of speech coders.

• Digital simultaneous voice and data (DSVD): Voice requires real-time transmission while data does not. By allocating the data bit stream between talk spurts, both voice and data can be transmitted simultaneously through a single channel.

3.3 Problems in VAD

VAD can be treated as a binary detection technique with the output of either speech presence or absence for a segment of input signals. The problem seems to be quite easy for clean speech. By measuring the energy level of the input signal, in noise free environment, it would be possible to achieve a high detection performance. However, in real environments, the input signal normally features noisy speech. Under heavy noise environments, the speech can be obscured by the noise. Especially the unvoiced sounds, which is important in speech intelligibility, might be mis-detected under noise environments. An example for a noisy speech segment with vehicle noise of 5 dB in signal to noise ratio (SNR) is shown in Figure 3.1. It shows that speech signals with low
energies become fully embedded in the noise, and does not seem to be easy to visually discriminate the talk spurs. Incorrect detection of talk spurs could generate clipping sounds, which could lead to severe degradation in voice quality. On the other hand, an increase in false detection of silences loses the potential benefits of silence compression. There is a trade-off in VAD, such that maximising the detection rate for speech while minimising the false alarm rate for silence.

3.4 Standard VAD Methods

This section reviews standards VAD algorithms together with performance comparison. In subsection 3.4.1, VAD algorithms are classified according to input features. Subsequently, brief characteristics of each VAD algorithm of ITU-T G.729B/G.723.1A, ETSI GSM-FR/HR/EFR, ETSI AMR, and TIA/EIA IS-127/733 are described in subsections 3.4.2, 3.4.3, 3.4.4, and 3.4.5, respectively. Experimental results together with discussions are presented in subsection 3.4.6.

3.4.1 Classification

In order to exploit silence compression, many VAD methods have been proposed, and some of them have been selected by standard organisations including the ITU-T, ETSI, and TIA/EIA. The ITU-T released G.729 Annex B (G.729B) [13] and G.723.1 Annex A (G.723.1A) [14] as extensions to 8-kbit/s G.729 [12] and 5.3/6.3-bit/s G.723.1 [10] speech coders for discontinuous transmission (DTX). The ETSI recommended GSM-FR, -HR, and -EFR VAD methods for European digital cellular systems [23][24][25]. Recently, the ETSI released another two VAD methods, adaptive multi-rate VAD option 1 (AMR1) and 2 (AMR2) [26], with a view to the third generation mobile communications for the UMTS. The North American standard organisation, TIA/EIA, released two VAD methods for IS-96 [27] and IS-127 [28] / IS-733 [29], in which the IS-127 and IS-733 VAD algorithms have a same structure.

Table 3.1 shows standard VAD techniques classified in terms of the input features mainly including sub-band energies and the spectral shape. The TIA/EIA VAD meth-
Table 3.1: Classification of standard VAD methods depending on input features. The values in parentheses indicate the number of spectral sub-bands.

<table>
<thead>
<tr>
<th>Main features</th>
<th>VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral shape</td>
<td>GSM-FR, GSM-HR, GSM-EFR</td>
</tr>
<tr>
<td>Sub-band energies</td>
<td>IS-96(1), IS-127(2), IS-733(2)</td>
</tr>
<tr>
<td></td>
<td>AMR1(9), AMR2(16)</td>
</tr>
<tr>
<td>Multi-boundary region classification</td>
<td>G.729B, G.723.1A</td>
</tr>
</tbody>
</table>

This table shows the classification of standard VAD methods depending on input features. The values in parentheses indicate the number of spectral sub-bands.

Methods exploits the energies of a relatively small number of sub-bands. IS-96 VAD does not decompose the band, but just use the overall signal energy. However IS-127 and IS-733 conduct VAD with two sub-band energies. Traditionally, ETSI VAD methods exploit the spectral shape. The motivation behind this is that the energy of the predictive coding error increases when the spectral shapes between the background and input signal mismatch. However, in the recent standard for AMR, they selected two VAD algorithms both of which are based on the spectral sub-band energy rather than the spectral shape. The ITU-T VAD standards, G.729B and G.723.1A, conduct the detection using four features including the spectral shape and sub-band energies. More detailed algorithmic characteristics of each standard VAD algorithm are described in the following few subsections.

3.4.2 ITU-T G.729B/G.723.1A VAD

As an extension to G.729 speech coder, the ITU-T SG16 released G.729 Annex B in order to support DTX by means of VAD, CNI, and CNG. G.729B conducts VAD decision every 10 ms, using four parameters:

- a full-band energy difference, \( \Delta E_f = \bar{E}_f - E_f \)
- a low-band energy difference, \( \Delta E_l = \bar{E}_l - E_l \)
3.4. Standard VAD Methods

- a spectral distortion, $\Delta LSF = \sum_{i=0}^{9}(LSF_i - LSF_i)^2$
- a zero-crossing rate difference, $\Delta ZC = ZC - ZC$

where $E_f$, $E_l$, $LSF_i$, and $ZC$ are the full-band energy, low-band energy, $i$-th line spectrum frequency, and zero-crossing rate of the input signal; and $\overline{E_f}$, $\overline{E_l}$, $\overline{LSF_i}$, and $\overline{ZC}$ are the noise parameters adapted from the background noise signal.

In specific, the input parameters for the VAD can be obtained from the input signal or from the intermediate values of the speech encoder. Subsequently, the difference parameters, $\Delta E_f$, $\Delta E_l$, $\Delta LSF$, and $\Delta ZC$ are computed from the input and noise parameters. A decision of voice activity is conducted over a four-dimensional hyperspace, based on a region classification technique, followed by a hangover scheme. The noise parameters are updated based on a first order auto-regressive (AR) scheme, if the full-band energy difference is less than a certain fixed threshold. The block diagram of G.729B VAD is shown in Figure 3.3. ITU-T G.723.1A VAD has a structure similar to G.729B VAD.

Figure 3.3: Block diagram of ITU-T G.729B VAD.
3.4. Standard VAD Methods

3.4.3 ETSI GSM-FR/HR/EFR VAD

The VAD algorithms for the ETSI GSM-FR-, -HR, and -EFR have a common structure, comparison of the prediction-residual energy with an adaptive threshold. The prediction-residual energy is computed using the current and smoothed autocorrelation ratios. The threshold for VAD decision is updated during noise-only regions using the most recent noise signals in order to reflect up-to-date noise characteristics. To be specific, the threshold adaptation is performed if the current frame is stationary but includes neither a pitch nor a tone. A frame is classified into stationary if the spectral distortion of the smoothed autocorrelation ratio between the current and previous frames is less than a threshold, otherwise nonstationary. The block diagram of GSM-EFR VAD is shown in Figure 3.4.

3.4.4 ETSI AMR VAD

AMR1 decomposes the input signal into nine sub-bands by means of the filter bank characterising a larger bandwidth for a higher frequency sub-band, and then calculates each sub-band energy followed by computing the SNR. The energy of the background
3.4. Standard VAD Methods

noise in the SNR is computed by an adaptive method based on a first order AR-model together with internal VAD logics. Finally, the voice activity is decided by comparing the sum of the sub-band SNRs with an adaptive threshold, followed by a hangover. The block diagram of AMR1 is shown in Figure 3.5.

AMR2 has a structure similar to AMR1 in that VAD is conducted using the sub-band energies together with the background noise energy. However, AMR2 transforms the input signal into the frequency domain using FFT, instead of the filter bank used in AMR1, and then calculates 16 sub-band energies with a nonlinear scale in band grouping. Subsequently, the SNR for each sub-band is calculated using the input and the background noise spectra. The background noise energy for each band is adapted during noise frames using a first-order AR-based scheme. AMR2 increases the threshold for final VAD decision for highly fluctuating signals, measured by the variance of instantaneous frame-to-frame SNRs, in order to prevent being over-sensitive to non-stationary background noise conditions. Additionally, noise adaptation may not be properly performed by measuring the spectral deviation when sub-band energies fluctuate high.
3.4. Standard VAD Methods

Thus, AMR2 changes the VAD threshold in an adaptive way together with the variation of burst and hangover counts. The hangover control is performed by measuring the peak-to-average SNR, in which the average SNR is calculated using AR-adaptation with the increased instantaneous SNR. In other words, for an increase of the peak-to-average SNR, it decreases the hangover and burst counts while increasing the VAD threshold. The block diagram of AMR2 is shown in Figure 3.6.

3.4.5 TIA/EIA IS-127/733 VAD

CDMA-based digital cellular systems have a natural structure for incorporating VAD, called a rate determination algorithm (RDA), which gives substantial improvement in channel capacity by controlling the radio transmission power to reduce cochannel interference. The TIA/EIA released two RDAs for IS-95 and IS-127, called 8-kbit/s Qual-
3.4. Standard VAD Methods

As input parameters, IS-127 RDA exploits two sub-band energies and the long-term prediction gain. Firstly, it calculates the smoothed sub-band energy using a first-order AR-model. Subsequently, the signal and noise energies for each sub-band are adapted depending on the long-term prediction gain. In other words, the signal energy is actively adapted to the current input if the prediction gain is relatively high. On the other hand, if the gain is relatively low, it increases the noise adaptation rate through by more weighting the input signal. Using the two sub-band energies of the signal and noise, each sub-band SNR can be calculated. The final rate is determined by comparing the SNRs with adaptive thresholds depending on the level of background noise and the SNR of the previous frame, followed by a hangover. The block diagram of IS-127 RDA is shown in Figure 3.7.
3.4. Standard VAD Methods

3.4.6 Experimental Results

The performances of the five standard VAD algorithms are evaluated in terms of speech and silence detection error rates. Speech materials of duration 96-sec are collected, filtered by the modified IRS, and then mixed with vehicle and babble noises of 5, 10, 15, and 25 dB SNR. The active and inactive regions of the speech material are marked manually. The proportions of the inactive and active regions of the speech material are 0.43 and 0.57, respectively. The VAD test for the processed noisy signal is carried out every 10 ms in the cases of G.729B and AMR2; but every 20 ms in GSM-EFR VAD, AMR1, and IS-127. With slight modification to the AMR2 source code, it is possible to obtain 10 ms results because AMR2 basically conducts the detection every 10 ms and then produces 20 ms results by a logical combination of the two 10-ms results. In handling the multiple rates of IS-127, the upper two rates, 1 and 1/2, and the lowest rate, 1/8, are treated as voice active and voice-inactive, respectively.

Performances under vehicle noise environment are shown in Figure 3.8 and 3.9, and those for babble noise are shown in Fig 3.10 and 3.11. G.729B exhibits the worst performances compared with other methods especially for low SNRs. G.729B produces enormous speech detection errors, which could cause severe clipping sounds of speech. IS-127 exhibits relatively high error rates for speech detection compared with those of ETSI VAD. However, it produces quite desirable performances in silence detection for babble noisy speech. The ETSI VAD methods, i.e. GSM-EFR VAD, AMR1, and AMR2, exhibit similar performances in speech detection while resulting in quite various performances in silence detection. GSM-EFR VAD produces the most desirable performances for relatively high SNRs, i.e. greater than 15 dB. However, the error rates of silence detection increase substantially for a decrease of the SNR. AMR2 produces relatively consistent results regardless of the noise levels in silence detection for vehicle noisy speech. The performance of AMR1 moderately locates between GSM-EFR VAD and AMR2. Comparative performances of the standard VAD algorithms are shown by examples in Figure 3.12, 3.13, 3.14, 3.15, 3.16, and 3.17 for various noise sources and levels.
Figure 3.8: Comparison of speech detection error rates against various vehicle noise levels.
3.4. Standard VAD Methods

Figure 3.9: Comparison of silence detection error rates against various vehicle noise levels.
Figure 3.10: Comparison of speech detection error rates against various babble noise levels.
Figure 3.11: Comparison of silence detection error rates against various babble noise levels.
3.4. Standard VAD Methods

Figure 3.12: Comparison by examples of VAD results over vehicle noise of 5 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.4. Standard VAD Methods

Figure 3.13: Comparison by examples of VAD results over vehicle noise of 15 dB SNR:
(a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.4. Standard VAD Methods

Figure 3.14: Comparison by examples of VAD results over vehicle noise of 25 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.4. Standard VAD Methods

Figure 3.15: Comparison by examples of VAD results over babble noise of 5 dB SNR:
(a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.4. Standard VAD Methods

Figure 3.16: Comparison by examples of VAD results over babble noise of 15 dB SNR:
(a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.4. Standard VAD Methods

Figure 3.17: Comparison by examples of VAD results over babble noise of 25 dB SNR: (a) noisy input speech, (b) clean speech, (c) G.729B, (d) IS-127, (e) GSM-EFR, (f) AMR1, and (g) AMR2.
3.5 VAD based on a Smoothed Statistical Likelihood Ratio

3.5.1 Motivation

Traditionally, VAD is performed on the basis of heuristics, thus it is not easy to analyse the characteristics of the detector and to improve the performance. Recently, Sohn et al. proposed a novel VAD method based on a statistical model, and reported that it can produce a high detection accuracy [38]. The reason for the high performance is attributed to the adoption of Ephraim and Malah’s noise suppression rules [39] for the voice activity decision rules, conducted by a likelihood ratio test using a decision-directed parameter estimator for an unknown parameter.

From investigation of Sohn et al.’s VAD, however, it is observed that the VAD may generate relatively high numbers of detection errors at the offset region of speech signals. Sohn et al. have circumvented this problem by a hangover scheme, but the reason for the undesirable phenomenon is not mentioned. We analysed the behavioural characteristics of Sohn et al.’s VAD, identified the rationale of the unwanted phenomenon, and then proposed a solution enabling significantly improved VAD.

3.5.2 Description of Decision Rules based on Likelihood Ratio Test

Voice activity decision can be considered as a test of two hypotheses: $H_0$ and $H_1$, which indicate speech absence and presence, respectively. Assuming that each spectral component of speech and noise has complex Gaussian distribution [39], in which the noise is additive and uncorrelated with the speech, the conditional probability density functions (PDF) of a noisy spectral component $Y_k$, given $H_{0,k}$ and $H_{1,k}$, are

$$p(Y_k|H_{0,k}) = \frac{1}{\pi \lambda_{N,k}} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{N,k}} \right\}$$

$$p(Y_k|H_{1,k}) = \frac{1}{\pi (\lambda_{N,k} + \lambda_{X,k})} \exp \left\{ -\frac{|Y_k|^2}{\lambda_{N,k} + \lambda_{X,k}} \right\}$$

where $k$ indicates the spectral bin index, and $\lambda_{N,k}$ and $\lambda_{X,k}$ denote the variances of the noise and speech spectra, respectively.
The likelihood ratio (LR) of the $k$th spectral bin, $\Lambda_k$, is defined from the above two PDFs as \[38\]

\[
\Lambda_k = \frac{p(Y_k|H_{1,k})}{p(Y_k|H_{0,k})} = \frac{1}{1 + \xi_k} \exp \left\{ \frac{(1 + \gamma_k)\xi_k}{1 + \xi_k} \right\} \tag{3.3}
\]

where $\gamma_k$ and $\xi_k$ are a posteriori and a priori SNRs, respectively, defined $\gamma_k = |Y_k|^2/\lambda_{N,k} - 1$ and $\xi_k = \lambda_{X,k}/\lambda_{N,k}$. Note that the definition of the a posteriori SNR is slightly different from the original one, $\gamma_k = |Y_k|^2/\lambda_{N,k}$ [39] for convenience in the later description. The noise variance is assumed to be known through noise adaptation (see Section 3.5.4). However, the variance of the speech is unknown, thus the a priori SNR of the $n$th frame, $\xi_k^{(n)}$, is estimated using the decision directed (DD) method [39] as

\[
\xi_k^{(n)} = \alpha \frac{|\hat{X}_k^{(n-1)}|^2}{\lambda_{N,k}^{(n-1)}} + (1 - \alpha)\text{MAX}\{\gamma_k^{(n)}, 0\} \tag{3.4}
\]

where $\alpha$ is the weighting term, e.g. 0.98, and the enhanced spectral amplitude $|\hat{X}_k|$ is estimated using the minimum mean square error of the log spectral amplitude estimator [40]. The decision of the voice activity is performed by the geometric mean of $\Lambda_k$ over all spectral bins as

\[
\Lambda = \exp \left\{ \frac{1}{K} \sum_{k=1}^{K} \log \Lambda_k \right\} \tag{3.5}
\]

where $K$ denotes the number of spectral bins.

The a posteriori SNR $\gamma_k$ fluctuates highly from frame to frame because of the high fluctuation of the short-time spectral amplitude $|Y_k|$. On the other hand, the estimated a priori SNR $\xi_k$ changes slowly due to the smoothing effect. As the value of $\alpha$ increases, so that of $\xi_k$ becomes smoother. The features of the two SNRs, $\gamma_k$ and $\xi_k$, compensate each other in the calculation of $\Lambda_k$, and consequently enable to enhance the performance of the VAD. The DD estimator for the a priori SNR is useful not only for avoiding the musical noise phenomenon in speech enhancement [41], but also for reducing the error rate in voice activity detection.

### 3.5.3 Analysis and Improvement of the Likelihood Ratio

The behavioural characteristics of the LR in (3.3) are observed with respect to the a priori and a posteriori SNRs, as shown in Figure 3.18. The maximal peaks in Figure
 Figure 3.18: Likelihood ratio versus a priori SNR versus a posteriori SNR. The solid lines from top-most to bottom are a posteriori SNRs of 15, 10, 5, 0, -5, -10, -15 dB, respectively.

3.18 correspond to the result of maximum likelihood (ML) estimation of the a priori SNR. The ML estimator [38] results in lower performance in comparison with the DD estimator because of the inherent high-fluctuation of the a posteriori SNR. The LR employing the DD estimator has the following properties:

1. If the a posteriori SNR is very high, i.e. $\gamma_k \gg 1$, and the range of the a priori SNR is limited properly, the LR becomes very high, i.e. $\Lambda_k \gg 1$.

2. If the a posteriori SNR is low, i.e. $\gamma_k < 1$, the a priori SNR becomes a key parameter in the LR.

In practice, the threshold of the LR is set to between 0.2 and 0.8 dB, and both the a posteriori and the a priori SNRs are bounded between -15 and 15 dB.

The delay of the noise variance $\lambda_{n,k}^{(n-1)}$ in (3.4) does not seriously affect the a priori SNR $\xi_k^{(n)}$, assuming that the noise statistics change slowly. However, the spectral amplitude
of the speech signal may change abruptly, particularly in onset and offset regions, in which the power of the spectral bins could rapidly increase and decrease, respectively. At the offset region, $\gamma_k$ can be low but $\hat{A}_k$ can be much higher than $\gamma_k$ due to the delay term $|\hat{X}_k^{(n-1)}|^2$ in (3.4). Thus $A_k$ becomes too low according to the LR property 2, and consequently it may become lower than the threshold of VAD. On the other hand, the delay rarely causes a problem at the onset regions, according to the LR property 1, as $\gamma_k^{(n)}$ in (3.3) is large enough.

A few techniques are investigated to overcome the problem in LR-based VAD. Firstly, it is possible to consider an adaptive weighting factor in the estimation of the *a priori* SNR in (3.4). In other words, a lower $\alpha$ can be assigned for the active region, but a higher $\alpha$ for the inactive region. When a low $\alpha$ is assigned at the offset region, it reduces the effect of the delay in (3.4), produces a lower $\hat{A}_k$, and therefore may prevent the abrupt decay of $A_k$. However, in our experiment, it was not easy to design a generalised adaptive rule which consistently gives satisfactory performance over various kinds of speech and noise signals. Thus, more investigation is required for attaining a generalised rule concerning the adaptive $\alpha$. Secondly, a smoothed likelihood ratio (SLR) $\Psi_k^{(n)}$ is considered and defined as

$$\Psi_k^{(n)} = \exp \left\{ \kappa \log \Psi_k^{(n-1)} + (1 - \kappa) \log A_k^{(n)} \right\}$$

where $\kappa$ is the smoothing factor, and $A_k^{(n)}$ is defined in (3.3) for the $n$th frame. The decision of the voice activity is finally carried out by

$$\Psi^{(n)} = \exp \left\{ \frac{1}{K} \sum_{k=1}^{K} \log \Psi_k^{(n)} \right\}$$

An $n$th input frame is classified as voice-active if $\Psi^{(n)}$ is greater than a threshold, and voice-inactive otherwise.

An example of the LR and the SLR over a segment of speech signals is shown in Figure 3.19(a), (b), and (c). The SLR seems to overcome the problem outlined for the LR. As shown in Figure 3.19(b), the SLR is relatively higher than the LR at the offset regions. The comparison over inactive frames is also shown in Figure 3.19(c), which indicates that the SLR fluctuates less than the LR.
3.5. VAD based on a Smoothed Statistical Likelihood Ratio

Figure 3.19: An example of the computed LR (solid line) and SLR (dotted line) of a segment of vehicle noisy signals of 5 dB SNR. The dotted horizontal-line indicates the VAD threshold. The boxed regions in (a) are enlarged in (b) and (c).
3.5. VAD based on a Smoothed Statistical Likelihood Ratio

### 3.5.4 Noise Estimation based on the SLR

The short-time spectral amplitudes of the noise signal could fluctuate severely from frame to frame, depending on the characteristics of the noise source. In order to mitigate this phenomenon, parameter smoothing techniques are considered in the estimation of the variance of noise spectra. Moreover, in order to cope with time-varying noise signals, the variance of the noise spectrum is adapted to the current input signal by a soft decision-based method.

The speech absence probability (SAP) of the kth spectral bin, \( p(H_{0,k}|Y_k) \), can be calculated by Bayes' rule as

\[
p(H_{0,k}|Y_k) = \frac{p(H_{0,k})p(Y_k|H_{0,k})}{p(H_{0,k})p(Y_k|H_{0,k}) + p(H_{1,k})p(Y_k|H_{1,k})}
\]

where \( p(H_{1,k}) = 1 - p(H_{0,k}) \), and the unknown a priori speech absence probability (PSAP), \( p(H_{0,k}) \), is estimated in an adaptive manner given as

\[
\hat{p}(H_{0,k}) = \min \{ \max \{ \beta \hat{p}(H_{0,k}^{(n-1)}) + (1 - \beta)p(H_{0,k}^{(n)}), H_0^{(L)}, H_0^{(U)} \} \}
\]

where \( \beta \) is the smoothing factor, e.g., 0.65. The lower and upper limits, \( H_0^{(L)} \) and \( H_0^{(U)} \), of the PSAP are determined through experiments, e.g., 0.2 and 0.8, respectively. Note that the SLR \( \Psi_k \) instead of the LR \( \Lambda_k \) is applied to the calculation of the SAP.

The variance of the noise spectrum of the kth spectral component in the nth frame, \( \lambda_{N,k}^{(n)} \), is updated in a recursive way as

\[
\lambda_{N,k}^{(n)} = \eta \lambda_{N,k}^{(n-1)} + (1 - \eta)E(\|N_k^{(n)}\|^2|Y_k^{(n)})
\]

where \( \eta \) is the smoothing factor, e.g., 0.95. The expected noise power-spectrum \( E(\|N_k^{(n)}\|^2|Y_k^{(n)}) \) is estimated by means of the soft-decision technique [42] as

\[
E(\|N_k^{(n)}\|^2|Y_k^{(n)}) = E(\|N_k^{(n)}\|^2|H_{0,k})p(H_{0,k}|Y_k^{(n)}) + E(\|N_k^{(n)}\|^2|H_{1,k})p(H_{1,k}|Y_k^{(n)})
\]

where \( p(H_{1,k}|Y_k^{(n)}) = 1 - p(H_{0,k}|Y_k^{(n)}) \). Through the experiments, it is observed that the SLR-based adaptation is useful for the estimation of the noise spectra with high fluctuation, such as a babble noise source.
3.5.5 Experimental Results

An objective test is conducted to evaluate the performance of the proposed VAD scheme. Speech materials of duration 96-sec are collected, filtered by the modified IRS, and then mixed with vehicle and babble noises of 5, 10, 15, and 25 dB SNR. The active and inactive regions of the speech material are marked manually. Furthermore in order to obtain more detailed information depending on the characteristics of speech signals, each active region is sub-classified into speech onset, speech offset, and strongly active speech (SAS) of which the proportions in the speech material are 0.43, 0.13, 0.17, and 0.27, respectively. The VAD test is carried out every 10-ms frame of the processed noisy signal.

The effect of the smoothing factor $\kappa$ in (3.6) is investigated, as shown in Figure 3.20. Note that the case of $\kappa = 0$ reduces (3.6) to the LR-based method. It is obvious that the detection accuracy, with an increase of $\kappa$, can be improved considerably at the offset region without serious degradation in the performance at the onset region for both vehicle and babble noisy signals. Concerning the detection of the inactive frames, interesting experimental results are observed. In the case of vehicle noisy signals, as $\kappa$ increases, the false alarm rate in the inactive frames increases gradually, for $\kappa < 0.9$, and then substantially, for $\kappa > 0.9$. However, in the case of babble noisy signals, it is observed that the error rate decreases gradually as $\kappa$ increases, for $\kappa < 0.9$, and then increases like the case of the vehicle noisy signal, for $\kappa > 0.9$. Therefore, if $\kappa$ is selected properly, the SLR-based method gives significantly improved performance compared with the LR-based method. The reason for this result is explained in Section 3.5.3.

Over various noise levels and sources, the performance of SLR-based VAD is compared with those of standard VAD, such as ITU-T G.729 annex B VAD (G.729B) [13] and ETSI AMR VAD option 2 (AMR2) [26], and LR-based VAD with and without the hangover scheme [38], as shown in Table 3.2. Original AMR2 conducts VAD every 10-ms, and then gives a 20-ms result. It decides the signal is voice-active if at least one of the 10-ms results is voice-active. Thus, the 10-ms VAD result can be obtained easily. Taking into account the results in Figure 3.20, $\kappa = 0.9$ is selected for SLR-based VAD. G.729B generates considerably high error rates at the active regions in comparison with
3.5. VAD based on a Smoothed Statistical Likelihood Ratio

Figure 3.20: Analysis of the smoothing factor $\kappa$ of the SLR with respect to detection error rates. The noise level is 10 dB SNR, and the noise sources are (a) vehicle and (b) babble.
other methods. It is important to note that frequent detection errors of speech frames lead to serious degradation in speech quality. Thus, emphasis should be put on reducing the error rate of speech detection rather than that of silence detection in alleviating the performance tradeoff between the speech and silence detections. LR-based VAD gives consistently superior performance to G.729B, but LR-based VAD without the hangover scheme produces relatively high detection error rates for voice-active regions. The hangover scheme can considerably alleviate this problem, but the speech detection error rate is still somewhat high in comparison with the results of both SLR-based VAD and AMR2. The performances between SLR-based VAD and AMR2 seem to be comparable.

3.6 Summary

In this chapter, we have reviewed standard VAD techniques and proposed SLR-based VAD. Through performance evaluation of the standard VAD methods, including G.729B, GSM-EFR VAD, AMR1, AMR2, and IS-127, it has been found that both AMR1 and AMR2 produce relatively high and consistent performances over various noise sources and levels.

Subsequently, we have analysed the behaviour of a VAD method based on the statistical likelihood ratio (LR), and found that the delay term in the decision-directed parameter estimator employed for the estimation of the LR can cause frequent detection errors in the offset regions of speech frames. In order to circumvent this problem, we have proposed the smoothed likelihood ratio (SLR), which provides a graceful decrease of the likelihood ratio at the offset region. Moreover, the SLR-based parameter smoothing technique is applied for adaptation of the noise variance so as to cope with high fluctuation of the noise spectra. Through the experiments, it has been shown that the proposed SLR scheme is highly desirable for the improvement of LR-based VAD. Additionally, SLR-based VAD gives detection performances superior to G.729B and comparable with AMR2.
Table 3.2: Comparison of speech and silence detection error rates among SLR-based, LR-based, AMR2, and G.729B VADs. The LR+HO means LR-based VAD with the hangover scheme.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>VAD</th>
<th>Detection error rate (%)</th>
<th>Vehicle noise</th>
<th>Babble noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Inactive</td>
<td>Onset</td>
<td>Offset</td>
</tr>
<tr>
<td>5</td>
<td>SLR</td>
<td>13.87</td>
<td>6.42</td>
<td>7.51</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>4.49</td>
<td>12.88</td>
<td>30.92</td>
</tr>
<tr>
<td></td>
<td>LR+HO</td>
<td>5.33</td>
<td>12.05</td>
<td>12.86</td>
</tr>
<tr>
<td></td>
<td>AMR2</td>
<td>18.64</td>
<td>9.13</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>G.729B</td>
<td>8.58</td>
<td>70.23</td>
<td>60.21</td>
</tr>
<tr>
<td>15</td>
<td>SLR</td>
<td>17.12</td>
<td>3.48</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>5.07</td>
<td>5.34</td>
<td>19.85</td>
</tr>
<tr>
<td></td>
<td>LR+HO</td>
<td>7.52</td>
<td>4.75</td>
<td>6.80</td>
</tr>
<tr>
<td></td>
<td>AMR2</td>
<td>20.15</td>
<td>3.78</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>G.729B</td>
<td>8.57</td>
<td>31.19</td>
<td>39.41</td>
</tr>
<tr>
<td>25</td>
<td>SLR</td>
<td>23.01</td>
<td>2.82</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>6.64</td>
<td>3.29</td>
<td>11.79</td>
</tr>
<tr>
<td></td>
<td>LR+HO</td>
<td>10.94</td>
<td>1.56</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>AMR2</td>
<td>20.28</td>
<td>2.68</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>G.729B</td>
<td>8.85</td>
<td>12.75</td>
<td>19.06</td>
</tr>
</tbody>
</table>
Chapter 4

Speech Enhancement

This chapter aims to achieve enhanced speech under noise background conditions. Section 4.1 describes the motivation and classification of speech enhancement techniques. Next, in Section 4.2, various kinds of STSA-based speech enhancement techniques are reviewed. Finally, in Section 4.3, a novel noise adaptation method is proposed.

4.1 Introduction

In voice communications, speech signals can be corrupted by environmental noise, and subsequently the communication could be fatiguing and less intelligible. Furthermore, compression of the noisy speech with a low bit-rate vocoder may lead the speech quality to considerable degradation due to frequent estimation errors of the model parameters constituting the vocoder. This problem can be treated by speech enhancement, which could present more pleasant voice communication by suppressing the noise components in input signals.

Let us assume that the noisy signal is formed additively by speech and noise signals in which the noise is generated by environmental sources such as vehicle, street, babble, etc. Basically, in real environments, complete noise cancellation is not feasible as it is not possible to completely track the noise characteristics which change with the time. However, by assuming that the noise characteristics change slowly in comparison with
the speech, it is possible to achieve a mitigated solution producing a more pleasant quality of voice after noise reduction. In speech coding, the noise signal could cause parameter estimation errors, and subsequently result in degrading the speech quality. Hence, speech enhancement is becoming a common approach as a preprocessor for recent speech coders.

Speech enhancement techniques can be classified, depending on the number of available microphones, into single and multiple channels. In the case of a single channel, the reference noise is not available explicitly. The noise statistics are typically characterised during voice-inactive regions between talk spurts using a voice activity detector. On the other hand, when dual channels are available, one microphone senses the noisy speech, but the other can be used mainly to catch the noise. By eliminating the noise factor collected by the latter microphone from the former one, it would be possible to cancel the noise more efficiently. However, in real environments, the multiple microphone scheme can be limited in its application. Thus in this thesis we consider the case of a single microphone only.

For the last three decades, many kinds of speech enhancement techniques have been proposed [43] [44] [45] [46], mostly based on transform domain techniques, adaptive filtering, and model-based methods. The transform-based technique transforms the time domain signal into other domains, suppresses noise components, and then applies the corresponding inverse transform to reconstruct enhanced speech signals. Discrete Fourier transform (DFT), discrete cosine transform (DCT), Karhunen-Loève transform (KLT), and wavelet transform (WT) are widely known transform methods. The DFT-based technique has been intensively investigated based on short-time spectral amplitudes (STSA). The KLT-based technique, called a signal subspace-based method [47], decomposes the space into signal (or speech) and noise subspaces by means of eigen decomposition, and then suppresses the noise component in the eigenvalues. The DCT-based technique [48][49] is based on the ideas of the suboptimality feature of KLT, low computational complexity, and higher frequency resolution compared with DFT-based methods. It is also possible to consider the WT-based method in order to simultaneously exploit the time and frequency characteristics of noisy speech signals. Secondly, adaptive filtering cancels the noise using adaptive filters such as the Kalman filter.
The Kalman filter models noisy speech signals in terms of state space and observation equations, which represent the speech production process and the noise addition model together with channel distortion, respectively [46]. Kalman filters normally assume white Gaussian noise distribution. However, Gibson et al. proposed a generalisation over coloured noise signals [50][51]. Lastly, the model-based technique classifies the noisy signal using a priori speech model, such as hidden Markov and voiced/unvoiced models, and then conducts the enhancement depending on classified speech models [44]. This method could be useful for obtaining improved noise reduction performance for various kinds of speech signals. However, it requires extra training to build the model with intensive computation. Even worse, it could exhibit model selection errors. Fundamentally, it is not easy to handle complicated speech signals with a finite number of speech models.

In the various speech enhancement techniques, DFT (or STSA)-based methods have been well investigated in the forms of spectral subtraction, Wiener filtering, maximum likelihood-STSA estimation, and minimum mean square error-STSA estimation. The reason for the popularity of the STSA-based speech enhancement is owing to not only its computation simplicity but also recent technical advances for speech quality improvement. In the following section, details on the STSA-based speech enhancement techniques are reviewed.

### 4.2 Review of STSA-based Speech Enhancement

This section reviews STSA-based speech enhancement algorithms with their performances. At first, subsection 4.2.1 describes the assumption, objective, and overall structure of STSA-based speech enhancement methods. Subsequently, subsections 4.2.2, 4.2.3, 4.2.4, 4.2.5, and 4.2.6 present the formulations of spectral estimation techniques based on spectral subtraction, maximum likelihood STSA, Wiener filtering, minimum mean square error STSA, and speech presence uncertainty, respectively. Finally, the performances of the STSA-based enhancement methods are evaluated and discussed in subsections 4.2.7 and 4.2.8, respectively.
4.2. Review of STSA-based Speech Enhancement

4.2.1 Problem Definition

Assuming that the noise \( d(n) \) is additive to the speech signal \( x(n) \), the noisy speech \( y(n) \) can be written as

\[
y(n) = x(n) + d(n), \text{ for } 0 \leq n \leq K - 1,
\]

where \( n \) is the time index. The objective of speech enhancement is to find optimal enhanced speech \( \hat{x}(n) \) given \( y(n) \) with the assumption that \( d(n) \) is uncorrelated with \( x(n) \).

The time-domain signals can be transformed to the frequency domain as

\[
y_k = X_k + D_k, \text{ for } 0 \leq k \leq K - 1,
\]

where \( Y_k, X_k, \) and \( D_k \) denote the short-time DFT results of \( y(n), x(n), \) and \( d(n) \), respectively.

The STSA-based speech enhancement filters out the noise using the spectral amplitudes of the variables in (4.2). The enhanced spectrum \( \hat{X}_k \) can be written in terms of the gain \( G_k \) and the noisy spectrum \( Y_k \) as

\[
\hat{X}_k = G_k Y_k
\]

where \( 0 \leq G_k \leq 1 \). The gain \( G_k \) is a function in terms of a posteriori SNR

\[
\gamma_k = \frac{|Y_k|^2}{E(|D_k|^2)}
\]

and a priori SNR

\[
\xi_k = \frac{E(|X_k|^2)}{E(|D_k|^2)},
\]

where \( E(|D_k|^2) \) and \( E(|X_k|^2) \) are the statistical variances of the \( k \)th spectral components of the noise and speech, respectively. The function definition of the gain \( G_k \) depends on specific enhancement methods. The a posteriori SNR \( \gamma_k \) in (4.4) can be obtained easily as \( Y_k \) is the input noisy spectrum and \( E(|D_k|^2) \) can be obtained through a noise adaptation procedure (see Section 4.3). However, the speech variance \( E(|X_k|^2) \)
for the estimation of $u_k$ in (4.5) is not available. As a solution, Ephraim and Malah proposed the decision-directed (DD) method [39], given

$$\hat{\xi}_k^{(t)} = \alpha \frac{\hat{X}_k^{(t-1)}}{E(|D_k^{(t)}|^2)} + (1 - \alpha) \text{MAX}(\gamma_k^{(t)} - 1, 0)$$  \hspace{1cm} (4.6)$$

where $0 \leq \alpha < 1$, and $t$ is the frame index.

The task of speech enhancement can be considered as an optimisation problem, minimisation of the residual noise while keeping the speech quality. The residual noise means the difference between the original and estimated speech signals. The optimisation typically exhibits trade-off. Over-estimation of the noise statistics may degrade the speech quality or intelligibility, while under-estimation could generate considerable residual noise sounds. The most typical residual noise in speech enhancement is the musical noise, or called tonal noise, which is composed of narrow-band signals appearing and disappearing with time varying amplitudes and frequencies.

The overall block diagram of the general STSA-based speech enhancement method is shown in Figure 4.1. The noisy speech $y(n)$ is first converted into the STSA $|Y_k|$ by the DFT with windowing. The enhanced spectral amplitude $|\hat{X}_k|$ is estimated by multiplying the spectral gain $G_k$ with $|Y_k|$. Enhanced speech $\hat{x}(n)$ can be constructed by applying the inverse DFT to the enhanced STSA $|\hat{X}_k|$ with the noisy speech phase $\angle Y_k$, followed by an appropriate overlap-and-add procedure to compensate for the window effect and to alleviate abrupt signal changes between two consecutive frames. The methods how to attain the gain $G_k$ can be found in the following subsections.

### 4.2.2 Spectral Subtraction

The noisy spectrum $Y_k$ in (4.2) can be converted into the power spectrum as

$$|Y_k|^2 = |X_k|^2 + |D_k|^2 + X_k^*D_k + X_kD_k^*$$  \hspace{1cm} (4.7)$$

where $X_k^*$ and $D_k^*$ denote the complex conjugates of $X_k$ and $D_k$, respectively. In order to estimate $|X_k|^2$, the statistical expectation is applied to (4.7) since $|D_k|^2$, $X_k^*D_k$, and $X_kD_k^*$ are not available. It gives

$$|Y_k|^2 = |\hat{X}_k|^2 + E(|D_k|^2) + E(X_k^*D_k) + E(X_kD_k^*)$$  \hspace{1cm} (4.8)$$
where $E(\cdot)$ is the ensemble average, and $|\tilde{X}_k|^2$ is the enhanced power spectrum. The expected noise $E(|D_k|^2)$ can be given by a noise adaptation procedure (see Section 4.3). Due to the assumption that $x(n)$ is uncorrelated with $d(n)$, $E(X_k^*D_k) = 0$ and $E(X_kD_k^*) = 0$. Thus, (4.8) can be rewritten as

$$|Y_k|^2 = |\tilde{X}_k|^2 + E(|D_k|^2). \quad (4.9)$$

The enhanced power spectrum $|\tilde{X}_k|^2$ can be estimated by subtracting $E(|D_k|^2)$ from $|Y_k|^2$, called power spectral subtraction.

The spectral power subtraction can be generalised with an arbitrary spectral order, called generalised spectral subtraction (GSS), as

$$|Y_k|^\nu = |\tilde{X}_k|^\nu + E(|D_k|^\nu) \quad (4.10)$$

where $\nu$ is the spectral order. In the cases of $\nu = 1$ and $\nu = 2$, GSS in (4.10) can be reduced to the magnitude and power spectral subtractions, respectively.

In practice, GSS-based speech enhancement typically exhibits severe musical noise sounds due to the high fluctuation of the STSA of noisy signals. In other words, the estimated noise magnitude can be larger than the input spectral magnitude. In this case, the enhanced spectral magnitudes are clamped to zero in order to prevent
the spectral magnitude from being negative. The clamping which happens irregularly with the frequency and time leads to producing the sound of musical tones.

Berouti et al. proposed a method for alleviating the musical noise phenomenon [52]. Let $|\tilde{X}_k|' = |Y_k|' - \alpha E(|D_k|')$, Berouti's GSS (GBSS) is given

$$|\tilde{X}_k|' = \begin{cases} |\tilde{X}_k|', & \text{if } |\tilde{X}_k|' > \beta E(|D_k|'), \\ \beta E(|D_k|'), & \text{otherwise,} \end{cases}$$

(4.11)

where $\alpha$ and $\beta$ are the spectral over-subtraction and floor factors, respectively, with $\alpha \geq 1$ and $0 \leq \beta \leq 1$. Note that GBSS is reduced to GSS if $\alpha = 1$ and $\beta = 0$, and to PSS if $\nu = 2$, $\alpha = 1$, and $\beta = 0$. GBSS is capable of reducing the overall residual noise level and a number of musical noises by calibrating $\alpha$ and $\beta$, respectively. The GBSS gain $G_k^{(GBSS)}$ becomes

$$G_k^{(GBSS)} = \begin{cases} \left[ 1 - \alpha \left( \frac{1}{\gamma_k} \right)^{\frac{1}{2}} \right]^{\frac{1}{\nu}} & \text{if } \gamma_k^{\frac{1}{\nu}} > \alpha + \beta, \\ \beta^{\frac{1}{\nu}} - \frac{1}{\sqrt{\gamma_k}} & \text{otherwise.} \end{cases}$$

(4.12)

The noise floor factor $\beta$ contributes to the reduction of musical noise sounds. It has the effect of converting the narrow-band musical noise into a wider band noise. The higher $\beta$ is, the less musical noise is produced. However, if too high $\beta$, it results in an increase of the level of other kinds of residual noise. The over-subtraction factor $\alpha$ is useful for reducing the overall level of residual noise. In other words, the higher $\alpha$ gives the lower level of the residual noise. However, if too high $\alpha$, distortion in speech can be perceived. Through experiments, it is found that GBSS with $\nu = 2$, $\alpha = 4 \sim 8$, and $\beta = 0.1$ gives a moderate level of musical noise reduction while keeping the speech quality.

In GBSS, both spectral over-subtraction and floor factors are fixed to constant values. Speech enhancement based on GBSS exhibits various noise reduction performances depending on the selection of these two factors. There are approaches to obtain the optimal factors based on the psycho-acoustic model [53] and a parametric formulation [54]. In the psycho-acoustic approach, both $\alpha$ and $\beta$ change each frame depending on the psychoacoustic masking threshold for each spectral component [53]. In the parametric formulation, $\alpha$ is derived using the MMSE-based metric [54].
4.2.3 Maximum-likelihood Spectral Amplitude Estimation

In DFT-based speech enhancement, given $Y_k = X_k + D_k$, the optimum estimate of the speech amplitude $|X_k|$ is desired for the noisy spectrum $Y_k$, in which $X_k = |X_k| \exp(j\theta_k)$, where $\theta_k$ is the phase of $X_k$. Assuming that the noise $D_k$ has complex Gaussian distribution, the probability density function (PDF) of $Y_k$ conditioned over $|X_k|$ and $\theta_k$ is

$$ p(Y_k|X_k, \theta_k) = \frac{1}{\pi E(|D_k|^2)} \exp \left\{ -\frac{|Y_k|^2 - 2|X_k| \Re(e^{-j\theta_k}Y_k) + |X_k|^2}{E(|D_k|^2)} \right\}. \tag{4.13} $$

The maximum likelihood (ML) estimate of $|X_k|$, maximising $p(Y_k|X_k, \theta_k)$, can be obtained from the derivative to the PDF with respect to $|X_k|$ [55]. Through proper derivation steps, the ML estimate $|\hat{X}_k|$ is given

$$ |\hat{X}_k| = \frac{1}{2}(|Y_k| + \sqrt{|Y_k|^2 - E(|D_k|^2)}), \tag{4.14} $$

which can be written in term of the gain as

$$ G_{k}^{(ML)} = \frac{1}{2} + \frac{1}{2} \sqrt{1 - \frac{1}{\gamma_k}}. \tag{4.15} $$

4.2.4 Wiener Filtering

The Wiener filter (WF) is a minimum mean square error (MMSE) estimate of a desired signal in the time domain [43][46]. Given a noisy signal $y(n)$, for $0 \leq n \leq N - 1$, the Wiener filter produces the MMSE estimate signal $\hat{x}(n)$ of the desired signal $x(n)$ as

$$ \begin{pmatrix} \hat{x}(0) \\ \hat{x}(1) \\ \vdots \\ \hat{x}(N-1) \end{pmatrix} = \begin{pmatrix} y(0) & y(-1) & \cdots & y(1-P) \\ y(1) & y(0) & \cdots & y(2-P) \\ \vdots & \vdots & \ddots & \vdots \\ y(N-1) & y(N-2) & \cdots & y(N-P) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{P-1} \end{pmatrix}, \tag{4.16} $$

where $w_k$ is the filter coefficients for $0 \leq k \leq P - 1$ with the filter order $P$. Eq. (4.16) can be rewritten in the algebraic form as

$$ \begin{pmatrix} \hat{x} \end{pmatrix} = \begin{pmatrix} Y \end{pmatrix} \begin{pmatrix} w \end{pmatrix}. \tag{4.17} $$
The Wiener filter error signal $e$ is the difference between the desired and estimated speech signals as

$$e = x - \hat{x}. \quad (4.18)$$

The error metric $\varepsilon$ is defined as

$$\varepsilon = e^{T}e$$

$$= (x - Yw)^{T}(x - Yw)$$

$$= x^{T}x - w^{T}Y^{T}x - x^{T}Yw - w^{T}Y^{T}Yw. \quad (4.19)$$

The filter coefficients $w$ are derived by setting the derivative of $\varepsilon$ to zero with respect to $w$, as

$$\frac{\partial \varepsilon}{\partial w} = -2(x^{T}Y - w^{T}Y^{T}Y) = 0. \quad (4.20)$$

Then, the optimal $w$ is given by

$$w = (Y^{T}Y)^{-1}Y^{T}x \quad (4.21)$$

in which $Y^{T}Y$ and $Y^{T}x$ mean the autocorrelation matrix $R_{yy}$ of $y(n)$ and the cross-correlation vector $r_{yz}$ between $y(n)$ and $x(n)$, respectively. Thus, (4.21) can be written as

$$w = R_{yy}^{-1}r_{yx}. \quad (4.22)$$

Note that $R_{yy} = R_{xx} + R_{dd}$ and $r_{yx} = r_{xx}$; because of the assumption that the speech signal is uncorrelated with the noise signal. Thus, (4.22) becomes

$$w = (R_{xx} + R_{dd})^{-1}r_{xx}. \quad (4.23)$$

Eq. (4.23) can be interpreted in the frequency domain as

$$G_{k}^{(WF)} = \frac{E(|X_{k}|^2)}{E(|X_{k}|^2) + E(|D_{k}|^2)}$$

$$= \frac{\xi_{k}}{1 + \xi_{k}}. \quad (4.24)$$

Note that the Wiener filter gain $G_{k}^{(WF)}$ in (4.24) is defined in terms of the a priori SNR $\xi_{k}$ only.
4.2.5 MMSE Spectral Amplitude Estimation

The Wiener filter is time-domain MMSE estimation while the McAulay's method is frequency domain ML estimation. Thus, it is possible to consider the MMSE estimate of the spectral amplitude [39] which minimises

$$\varepsilon = (|X_k| - |\hat{X}_k|)^2. \quad (4.25)$$

The MMSE-STSA estimate $|\hat{X}_k|$, given $Y_k$, is

$$|\hat{X}_k| = \frac{1}{\pi B(|D_k|^2)} \exp \left\{ - \frac{|Y_k - \alpha_k e^{j\theta_k}|^2}{E(|D_k|^2)} \right\} \quad (4.27)$$

$$\frac{1}{\pi B(|D_k|^2)} \int_0^\infty \int_0^{2\pi} \alpha_k \rho(Y_k|\alpha_k, \theta_k) d\theta_k d\alpha_k$$

$$\int_0^\infty \int_0^{2\pi} \rho(Y_k|\alpha_k, \theta_k) d\theta_k d\alpha_k$$

$$p(Y_k|\alpha_k, \theta_k) = \frac{1}{\pi B(|X_k|^2)} \exp \left\{ - \frac{\alpha_k^2}{E(|X_k|^2)} \right\}$$

$$p(\alpha_k, \theta_k) = \frac{\alpha_k}{\pi B(|X_k|^2)} \exp \left\{ - \frac{\alpha_k^2}{E(|X_k|^2)} \right\}$$

$$p(\alpha_k) = \frac{2\alpha_k}{E(|X_k|^2)} \exp \left\{ - \frac{\alpha_k^2}{E(|X_k|^2)} \right\}$$

$$p(\theta_k) = \frac{1}{2\pi}$$

Through proper derivation steps [39], (4.26) can be rewritten as

$$|\hat{X}_k| = \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} \exp \left( - \frac{v_k}{2} \right) \left\{ \left( 1 + \frac{v_k}{2} \right) I_0 \left( \frac{v_k}{2} \right) + v_k I_1 \left( \frac{v_k}{2} \right) \right\} |Y_k| \quad (4.31)$$

where $\Gamma(\cdot)$ is the gamma function with $\Gamma(1.5) = \sqrt{\pi}/2$; $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively; and $v_k = \frac{\xi_k}{1+\xi_k} \gamma_k$.

As a variant, Ephraim and Malah proposed an MMSE log spectral amplitude (MMSE-LSA) estimator [40], based on the well known fact that a distortion measure with the log
spectral amplitudes is more suitable for speech processing. The MMSE-LSA estimator minimises the following distortion measure,

\[ \varepsilon = \left( \log |X_k| - \log |\hat{X}_k| \right)^2 \]  

with

\[ |\hat{X}_k| = \exp \left[ E \{ \log(|X_k|) | Y_k \} \right]. \]  

Through proper derivation steps [40], the final estimate becomes

\[ |\hat{X}_k| = \frac{\bar{x}_k}{1 + \bar{x}_k} \exp \left\{ \frac{1}{2} \int_{\bar{v}_k}^{\infty} e^{-t} dt \right\} |Y_k|. \]  

### 4.2.6 Spectral Estimation based on the Uncertainty of Speech Presence

The conventional speech enhancement methods can be extended by incorporating the uncertainty of speech presence [55][40]. The presence and absence of speech, \( H_0 \) and \( H_1 \), respectively, can be written as

\[ H_0 : \quad Y_k = D_k \]  
\[ H_1 : \quad Y_k = X_k + D_k. \]

Let assume that each spectral component of the speech and noise has complex Gaussian distribution, and the noise is additive to and uncorrelated with the speech signal. The conditional probability density functions observing a noisy spectral component \( Y_k \), given \( H_0 \) and \( H_1 \), are

\[ p(Y_k | H_0) = \frac{1}{\pi E(|D_k|^2)} \exp \left\{ - \frac{|Y_k|^2}{E(|D_k|^2)} \right\} \]  
\[ p(Y_k | H_1) = \frac{1}{\pi (E(|D_k|^2) + E(|X_k|^2))} \exp \left\{ - \frac{|Y_k|^2}{E(|D_k|^2) + E(|X_k|^2)} \right\} \]

where \( k \) is the spectral bin index, \( 0 \leq k \leq K/2 \); and \( E(|D_k|^2) \) and \( E(|X_k|^2) \) denote the variances of the \( k \)th spectral components of the noise and speech, respectively.

The probability of speech absence can be given by Bayes' rule as

\[ p(H_1 | Y_k) = \frac{p(Y_k | H_1) p(H_1)}{p(Y_k | H_0) p(H_0) + p(Y_k | H_1) p(H_1)} \]  
\[ = \frac{1}{1 + \varepsilon \Lambda_k} \]
where

\[ \varepsilon = \frac{p(H_1)}{p(H_0)} \]  

(4.40)

in which \( p(H_1) \) and \( p(H_0) \) denote the a priori probability of speech presence and absence, respectively. The likelihood ratio of the \( k \)th spectral bin \( \Lambda_k \) can be defined from the above two likelihoods, as

\[ \Lambda_k = \frac{p(Y_k|H_1)}{p(Y_k|H_0)} \]

\[ = \frac{1}{1 + \xi_k} \exp \left\{ \frac{(1 + \gamma_k)\xi_k}{1 + \xi_k} \right\} . \]  

(4.41)

The enhanced spectrum based on the probability of speech presence is written as

\[ \hat{X}_k = E(X_k|Y_k, H_0)p(H_0|Y_k) + E(X_k|Y_k, H_1)p(H_1|Y_k) \]  

(4.42)

where \( p(H_0|Y_k) \) denotes the probability of speech absence given \( Y_k \). Eq. (4.42) can be simplified as

\[ \hat{X}_k = E(X_k|Y_k, H_1)p(H_1|Y_k) \]  

(4.43)

since the expected speech spectrum under speech absence is null, i.e. \( E(X_k|Y_k, H_0) = 0 \). The first and second terms, \( E(X_k|Y_k, H_1) \) and \( p(H_1|Y_k) \), of the right hand side in (4.43) can be attained by a conventional spectral estimator and (4.39), respectively.

4.2.7 Experimental Results

Objective speech qualities for voice-active regions are evaluated in terms of both segmental SNR (SEGSNR) improvement and Itakura-Saito distortion (ISD). The SEGSNR improvement means the difference between the SEGSNRs of the enhanced speech and the noisy input signals, in which the SEGSNR is defined

\[ \text{SEGSNR}(\text{dB}) = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \left\{ \sum_{n=mM}^{(m+1)M-1} \frac{x^2(n)}{(x(n) - \hat{x}(n))^2} \right\} \]  

(4.44)

where \( N \) and \( M \) are the frame size and the total number of frames, respectively. The ISD is defined

\[ \text{ISD}(\text{dB}) = 10 \log_{10} \left\{ \frac{a_n^T R_{aa} a_n}{a_n^T R_{ab} a_n} \right\} \]  

(4.45)
4.2. Review of STSA-based Speech Enhancement

where \( a_d \) and \( a_e \) are the LPC coefficients of the desired and estimated speech signals, respectively; and \( R_e \) is the autocorrelation matrix of the estimated signal.

For the experiment, speech materials of 64 sec were collected, sampled at 8 kHz, mixed with the combinations of vehicle and helicopter noises of 0, 5 and 10 dB SNR, and then processed every 10 ms in the frequency domain by the five kinds of spectral estimators including PSS, GNSS, ML, WF, and MMSA-LSA. The MMSE-LSA is further classified, depending on the adoption of the speech presence uncertainty, into MMSE-LSA-HD and MMSE-LSA-SD in which HD and SD denote the hard and soft decision methods, respectively. The ideal signal, limited by the theory, is obtained using the original spectral amplitude with the phase of the noisy signal, because the speech enhancement is conducted with the enhanced spectral amplitudes and the phases of the noisy input speech.

The SEGSNR improvement and ISD for the vehicle and the helicopter noisy signals are shown in Figure 4.2, 4.3, 4.4, and 4.5. From the analysis, it is found that both WF and MMSE-based methods give satisfactory results compared to other methods.

For noisy input signals in Figure 4.6, the spectrograms of various enhancement methods are shown in Figure 4.9, 4.10, 4.11, 4.12, 4.13, and 4.14 in order to see the characteristics of the residual noise. The spectrograms of the noise-free and theoretical limit signals are also shown in Figure 4.7 and 4.8, respectively. For the PSS and ML-based methods, severe musical noise exhibiting irregular spots in the spectrograms can be observed in Figure 4.9 and 4.11, respectively. The GBSS method with \( \nu = 2, \alpha = 4, \) and \( \beta = 0.1 \) alleviates the vulnerable musical tones to a moderate level, as shown in Figure 4.10, compared with the PSS and ML methods. The WF-based method gives further reduction in the level of the residual noise as shown in Figure 4.12. Using the MMSE-STSA-based method, it is possible to even further eliminate the musical noises, as shown in Figure 4.13. Even though the level of the overall residual noise of the MMSE-STSA is slightly higher than that of the WF method, the sound quality of MMSE-STSA is perceptually more comfortable than that of the WF method. The higher speech-quality is owing to further reduction in tonal signals. Combining the soft-decision technique with the MMSE-based method, it is possible to even reduce the overall level of the
residual noise as shown in Figure 4.14.

4.2.8 Discussions

The Ephraim and Malah's speech enhancement method gives a high performance mainly due to the DD-based a priori SNR estimation [41]. Cappe [41] has shown its usefulness for eliminating musical noise phenomenon through behavioural analysis. From interpretation of (4.6), it is not difficult to see that $\hat{\xi}_k$ is a smoothed version of $\gamma_k$. The a posteriori SNR $\gamma_k$ features high fluctuation from frame to frame, while $\hat{\xi}_k$ changes slowly. By exploiting the characteristics of the two SNRs, $\gamma_k$ and $\hat{\xi}_k$, the high performance in speech quality is achieved.

The WF produces performances better than both GNSS and ML-based methods. The reason behind the high performance is also due to the DD-based a priori SNR in the gain function of the WF. The usefulness of the DD-based a priori SNR can be also applied to the a posteriori SNR-based speech enhancement methods, such as the GNSS and ML-based spectral estimators, by replacing the a posteriori SNR with the a priori SNR [56] as

$$\gamma_k = \hat{\xi}_k + 1.$$  \hspace{1cm} (4.46)

Although substantial alleviation of the musical noise phenomenon is achieved by the WF-based method, it is observed that the musical noise is not completely removed as shown in the spectrogram analysis. It is also possible to show that this phenomenon exists in the a priori SNR-based speech enhancement using (4.46). From investigation of the speech enhancement algorithms together with the speech quality analysis, we conclude with the following guidelines for developing a speech enhancement system with a high quality.

- Proper combination of the a priori and a posteriori SNRs is important to eliminate the musical noise while keeping the speech quality.

- The soft-decision technique based on the speech presence uncertainty is useful for further suppressing the level of the residual noise for voice-inactive regions.
4.2. Review of STSA-based Speech Enhancement

Figure 4.2: Comparison of SEGSNR improvements of STSA-based speech enhancement methods over vehicle noise environments.
4.2. Review of STSA-based Speech Enhancement

Theoretical limit — Unprocessed

Figure 4.3: Comparison of ISDs of STSA-based speech enhancement methods over vehicle noise environments.
4.2. Review of STSA-based Speech Enhancement

Figure 4.4: Comparison of SEGSNR improvements of STSA-based speech enhancement methods over helicopter noise environments.
4.2. Review of STSA-based Speech Enhancement

Figure 4.5: Comparison of ISDs of STSA-based speech enhancement methods over helicopter noise environments.
Figure 4.6: Noisy speech (a) waveform and (b) spectrogram. The noise source and level are vehicle and 5 dB SNR, respectively.
4.2. Review of STSA-based Speech Enhancement

Figure 4.7: Noise-free speech (a) Waveform and (b) spectrogram.
4.2. Review of STSA-based Speech Enhancement

Figure 4.8: Enhanced speech (a) Waveform and (b) spectrogram by theoretical limit.
4.2. Review of STSA-based Speech Enhancement

Figure 4.9: Enhanced speech (a) Waveform and (b) spectrogram by PSS.
4.2. Review of STSA-based Speech Enhancement

Figure 4.10: Enhanced speech (a) Waveform and (b) spectrogram by GBSS with $\nu = 2$, $\alpha = 4$, and $\beta = 0.1$. 
Figure 4.11: Enhanced speech (a) Waveform and (b) spectrogram by ML-STSA estimation.
4.2. Review of STSA-based Speech Enhancement

Figure 4.12: Enhanced speech (a) Waveform and (b) spectrogram by WF.
4.2. Review of STSA-based Speech Enhancement

Figure 4.13: Enhanced speech (a) Waveform and (b) spectrogram by MMSE-STSA estimation.
Figure 4.14: Enhanced speech (a) Waveform and (b) spectrogram by MMSE-STSA estimation with speech presence uncertainty.
4.3 Mixed Decision-based Noise Adaptation

Frequency domain speech enhancement focuses mainly on improved estimation of spectral attenuation factors with the assumption of given noise statistics. However, in practice, the noise statistics exhibit fluctuation from frame to frame. Thus, a method for robust estimation of the noise statistics is investigated.

Conventional noise estimation can be classified into two techniques. One is the hard decision (HD)-based method, which adapts the noise variance during voice-inactive regions by voice activity detection (VAD). This method is quite successful when VAD classifies the regions well. However, VAD itself is a complicated technique to implement when high performances under various noise sources and levels are required. Thus, speech detection errors due to VAD may cause over-estimation or under-estimation of the noise statistics, and lead to degradation of the speech quality. In other words, the performance of the HD-based method is subject to that of VAD.

The other technique is the soft decision (SD)-based method, which adapts the noise statistics based on the uncertainty of speech absence instead of the hard-limited function used in the HD-based method \[42]\[57]. This method does not rely on VAD decisions. However, it updates the noise statistics even in the presence of speech, and it is quite difficult to accurately measure the mixture ratio between speech and noise. The inaccurate measurement of speech absence, especially in voice-active regions, could seriously distort the enhanced speech.

Hence, we propose a novel noise adaptation method, named mixed decision (MD)-based noise adaptation, by taking into account the characteristics of the HD- and SD-based methods.

4.3.1 Hard Decision-based Noise Adaptation

The HD-based method conducts noise adaptation during speech absence regions only as

\[
E(|D_k^{(t)}|^2) = \begin{cases} 
\eta E(|D_k^{(t-1)}|^2) + (1 - \eta)|y_k^{(t)}|^2 & \text{if } y_k^{(t)} \in H_0 \\
E(|D_k^{(t-1)}|^2) & \text{otherwise}
\end{cases} \quad (4.47)
\]
4.3. Mixed Decision-based Noise Adaptation

where the superscript $t$ indicates the frame index, $\eta$ is the forgetting factor, e.g. 0.95, and $Y$ is the noisy spectrum. In the case of speech presence, detected by VAD, it does not update the noise variance. The HD-based noise adaptation has been widely used in speech enhancement.

4.3.2 Soft Decision-based Noise Adaptation

The SD-based noise estimation, the estimated noise given $Y_k$, is formulated as

$$E(D_k | Y_k) = \frac{E(D_k | Y_k, H_0)p(H_0 | Y_k) + E(D_k | Y_k, H_1)p(H_1 | Y_k)}{p(H_0 | Y_k) + p(H_1 | Y_k)G_{D,k} Y_k}$$

(4.48)

where $E(D_k | Y_k, H_0) = Y_k$, $E(D_k | Y_k, H_1) = G_{D,k} Y_k$. The probability of speech presence $p(H_1 | Y_k)$ is defined in (4.39), and $p(H_0 | Y_k) = 1 - p(H_1 | Y_k)$.

The optimal noise gain $G_{N,k}$ can be derived by the Wiener estimator $W$ in the time domain as designing a Wiener filter for the speech signal. It can be written, from proper derivation steps, as $W = (R_{xx} + R_{dd})^{-1} R_{dd}$ in which $R_{dd}$ and $R_{xx}$ denote the covariance matrix of the noise and speech signals. The frequency response of the derived filter becomes

$$G_{D,k} = \frac{E(|D_k|^2)}{E(|X_k|^2) + E(|D_k|^2)}$$

$$= \frac{1}{1 + \xi_k}$$

(4.49)

where $\xi_k$ is the a priori SNR which can be estimated using the decision-directed method defined in (4.6). Here, the noise gain $G_D$ can be separated as an independent task of estimation. Thus, it is also possible to estimate the noise gain using other kinds of enhanced spectral estimation techniques, such as MMSE, MMSE-LSA, etc. Note that the sum of $G_D$ and the speech gain $G$ is not necessarily 1.0, and the gain estimators applied to the speech and noise are different. The noise variance of the SD-based method can be estimated in a recursive manner as

$$E(|D_k(t)|^2) = \eta E(|D_k(t-1)|^2) + (1 - \eta) |E(D_k(t) | Y_k(t))|^2.$$  

(4.50)
4.3.3 Mixed Decision-based Noise Adaptation

In order to alleviate the problems in the HD- and SD-based methods, the MD-based method is proposed for noise adaptation as

\[
E(|D_k^{(t)}|^2) = \begin{cases} 
\eta E(|D_k^{(t-1)}|^2) + (1 - \eta)|Y_k^{(t)}|^2 & \text{if } Y^{(t)} \in H_0 \text{ and } \Lambda^{(t)} \leq \theta \\
\eta E(|D_k^{(t-1)}|^2) + (1 - \eta)E(D_k^{(t)}|Y_k^{(t)})^2 & \text{if } Y^{(t)} \in H_0 \text{ and } \Lambda^{(t)} > \theta \\
E(|D_k^{(t-1)}|^2) & \text{otherwise}
\end{cases}
\]

where \(\Lambda^{(t)} = \prod_{k=1}^{T_k} \Lambda_k^{(t)}\) defined in (4.41). The threshold \(\theta\) is set to a sufficiently small value that rarely classifies the speech into silence, i.e. \(\theta < 1\).

4.3.4 Experimental Results

Objective speech qualities are evaluated in terms of both SEGSNR improvement and ISD with respect to the speech detection error-rate of VAD (\(E_d\)), in order to investigate the robustness of the noise adaptation techniques. Various \(E_d\)s are calibrated by a voice activity detector [58], and then frame-by-frame VAD results are given to each noise adaptation method. For the experiment, speech materials of 64 sec were collected, sampled at 8 kHz, mixed with vehicle noise of 5 dB in SNR, and then processed every 10 ms in the frequency domain by the MMSE estimator [39] employing the noise adaptation methods. Finally, the enhanced speech signal is obtained by the inverse DFT of the enhanced spectrum followed by the overlap-and-add procedure.

Both SEGSNR improvement and ISD between the clean and enhanced speech signals for vehicle noisy speech signals of 0, 5, and 10 dB SNR are shown in Figure 4.15, 4.16, and 4.17, respectively. The experiments confirm that 1) the SD-based method results in lower performances compared to both the MD- and the HD-based method for low \(E_d\); 2) the HD-based method exhibits significant degradation in performance for an increase of \(E_d\); and 3) the MD-based method produces, regardless of the VAD performance, robust and superior performances in comparison with the HD- and SD-based methods. Note that for very low \(E_d\)s, i.e. \(0.0 \leq E_d < 0.1\), the performances of the MD and HD are slightly degraded compared with the case of \(E_d = 0.2\). This is caused by less frequent adaptation of the noise frames because of the increased false
alarm rate of the VAD. In other words, VAD produces the low $E_d$ at the expense of the increased false alarm rate during pauses.

Experimental results for helicopter noisy speech with levels of 0, 5, and 10 dB SNR are also shown in Figure 4.18, 4.19, and 4.20, respectively, which exhibit performance patterns similar to the vehicle noisy signals despite differences in the absolute values being measured.

## 4.4. Summary

This chapter has reviewed STSA-based spectral enhancement techniques including GSS, GBSS, ML, WF, and MMSE-based algorithms together with the estimate of speech presence uncertainty. Through intensive experiments, it has been found that MMSE-based STSA method combined with speech presence uncertainty is useful for noise reduction.

Noise adaptation in speech enhancement is mandatory to keep track of the noise characteristics. Thus, we have proposed the MD-based noise adaptation method for the robustness against the VAD errors, and have proven its usefulness through experiments.
Figure 4.15: Performance comparison in terms of SEGSNR improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 0 dB SNR.
Figure 4.16: Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 5 dB SNR.
Figure 4.17: Performance comparison in terms of the SEGSRN improvement and ISD against the speech detection error-rate of VAD for vehicle noisy speech of 10 dB SNR.
4.4. Summary

Figure 4.18: Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 0 dB SNR.
Figure 4.19: Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 5 dB SNR.
Figure 4.20: Performance comparison in terms of the SEGSNR improvement and ISD against the speech detection error-rate of VAD for helicopter noisy speech of 10 dB SNR.
Chapter 5

Speech Model Parameter
Estimation and Quantisation

This chapter aims to achieve enhanced algorithms for low bit-rate speech coding. Section 5.1 introduces a speech production model. Next, in Section 5.2, the redundancy in speech signals is discussed. Section 5.3 reviews classical low bit-rate speech coders. Section 5.4 presents a sinusoidal speech model together with conventional techniques for the estimation and the quantisation of the model parameters. Subsequently, Section 5.5 proposes a novel pitch estimation method. Section 5.6 and 5.7 propose efficient spectral magnitude quantisation methods for sinusoidal speech coders. Finally, in Section 5.8, the summary of the chapter is presented.

5.1 A Speech Production Model

Speech is a sequence of air pressure changes produced by the human vocal organs [59]. The lungs pump out an air flow through the vocal cord at the glottis, and then it passes through the vocal tract. The voiced sounds are produced by opening and closing the glottis. The unvoiced sounds are generated, with the glottis open, by forming constrictions at a high velocity to produce a turbulence. The sounds passing through the vocal tract forms resonance, and its frequencies are called formants. The change
in the vocal shape constructs speech information, which typically conveys linguistic information for communications among human beings.

Based on the physiological phenomenon, speech can be modelled by the source-filter model [59], meaning that speech is generated by exciting a time-varying filter with quasi-periodic and random pulses for voiced and unvoiced sounds, respectively. The filter and pulses simulate the vocal tract and source, respectively. Furthermore, the quasi-periodic and random pulses simulate the vibration and turbulence of the vocal source respectively.

Speech signals \( s(n) \) can be represented with convolution between the excitation source \( e(n) \) and the time-varying filter \( h(n) \) as

\[
    s(n) = e(n) * h(n)
\]  

(5.1)

where \( n \) is the sample index, and \( * \) is the convolution operator. The time domain formula (5.1) can be represented in the frequency domain as

\[
    S(f) = E(f)H(f)
\]  

(5.2)

where \( S(f) \), \( E(f) \), and \( H(f) \) are the DFT results of \( s(n) \), \( e(n) \), and \( h(n) \), respectively.

### 5.2 Redundancy in Speech

Speech coding techniques investigate redundancy features based on statistic and psychoacoustic characteristics of speech, and then apply them for efficient compression.

The statistical redundancy can be exploited by decomposing the speech into the source and the filter as (5.1). Note that, if a signal can be decomposed into independent components, the sum of the entropies of each separated component is less than the entropy of the signal itself according to the entropy coding theory. Typically, we model the filter using linear prediction (LP) coefficients based on the auto-regressive model. The poles of the filter correspond to the formants of speech. Speech decomposition into LP-coefficient and the residual signal can be seen at the point of short-term prediction, based on the fact that speech exhibits high correlation between neighbouring signals.
Additionally, it is not difficult to see that periodic components for voiced speech are formed by the periodic excitation source. By exploiting the pitch periodicity, it is possible to further decompose the residual signal into a long-term predictive and the secondary residual terms [60]. Hence, it is possible to achieve an improved speech coding gain from the decomposed speech signals, composed of the short-term correlation term, long-term correlation term, and secondary residual signal.

Investigation of psychophysical redundancy in speech is fundamental in designing low bit-rate speech coders because the human auditory system (HAS) can not perceive all the features of the sounds. Thus, the compression gain can be also achieved by removing the imperceptible components in speech. Masking of the HAS is a well-known psychoacoustic phenomenon. When there is a strong masker in time domain signals, weak signals before and after the masker may not be audible. The masking phenomenon also exists in the frequency domain. A strong frequency component masks other frequency components around the masker, called simultaneous masking. Depending on the power of frequency components, it is possible to design a perceptually improved quantiser. In speech, the formants exhibit a relatively high power compared with the spectral null parts. Thus, the formants can be emphasised too much in quantisation of the speech. The low energy parts, typically the spectral null parts and high frequency components, can be relatively neglected. However, those low energy components are perceptually important especially for nasal sounds. To alleviate this problem, it is quite general that low bit-rate speech coders introduce the perceptual weighting filter [60] in quantisation of the residual signal.

Another important auditory feature is the phase term of speech signals. The HAS is rather insensitive to phase changes of speech signals. In specific, for the linear phase changes, the HAS cannot perceive the distortion in signals. Hence, enhanced speech compression could be attained by exploiting the feature of the phase term of speech.

5.3 Classical Low Bit-rate Speech Coders

Development of low bit-rate speech coders, typically operating 1 ~ 8 kbit/s, has been challenged by many researchers with proposals of various kinds of speech coders. Even
though the approaches for various low bit-rate coders are different, they have a common basic structure: designed on the basis of the source-filter model. The filter in (5.1) can be represented by the LP coefficients. The residual signal, the output of the LP-inverse filter, becomes the excitation signal of the filter. Speech coders which have been proposed so far mainly investigate techniques how to efficiently represent the residual signal because of its crucial role in forming baseline speech-quality. Thus, from investigation of representation schemes of the residual signal, it is possible to classify and characterise speech coders.

The linear predictive coding (LPC)- vocoder represents the residual signal using the pitch and the voicing flag. The pitch represents the quasi-periodic pulses with a primary periodic-period for voiced speech. The voicing flag selects an excitation source type, periodic or random. The LPC vocoder can be designed around 2 kbit/s. It is, however, limited to communication quality because the model is too simple to represent complicated speech signals. Even worse, in the case of estimation errors of the model parameters including the voicing and pitch, the speech quality could be degraded severely.

Many complicated techniques exploiting the quasi-periodic nature of the excitation signals have been proposed. Regular pulse-excited (RPE)-LPC [61], the ETSI GSM-FR standard coder, represents them using regular-spaced pulses in which the information required for transmission is the start position and the width of the regular pulses. Multipulse-excited (MPE) coding [62] [63] represents the residual signal with multiple pulses. In other words, each pulse position with the gain of the multiple pulses is required for the representation. Actually, it does not conduct pitch period determination, but it detects a certain number of optimum or suboptimum pulses for a segment of speech signals by means of an analysis-by-synthesis technique [62]. The RPE-LPC and MPE coding techniques present much better quality speech compared to the LPC vocoder at the sacrifice of increased bit-rates. Code-excitation linear prediction (CELP)-based coders [64] quantise the excitation signal with pre-stored random or pulse code vectors, which typically gives further improved rate-distortion performance. CELP coders produce toll or near-toll quality speech with bit rates around 4 ~ 16 kbit/s. The high speech-quality of CELP coders is mainly due to 1) analysis-
by-synthesis (AbS) -based code excitation search [62], 2) the perceptual weighting filter (PWF) [60] [62], and 3) long-term prediction (LTP) [60]. The PWF exploits a perceptual redundancy in speech by emphasising the spectral nulls rather than the peaks in the process of AbS-based code vector search [65] [62]. LTP exploits long-term correlation, formed by the quasi-periodic excitation source, of speech signals. CELP coders present relatively consistent speech-quality compared to model-based coders, owing to the non-parametric approach for the representation of the residual signal. The performance is not sensitive to the misdetection of the pitch. Furthermore, the voicing decision is not necessary in designing a baseline CELP coder. For a decrease in the bit-rate lower than 4 kbit/s, however, CELP coders seems to give considerable degradation in speech quality. The CELP coding scheme still has the feature of waveform coders, representing the waveform shape. In other words, the phase redundancy is not exploited.

Sinusoidal speech coders [66][67][68], called harmonic speech coders or multiband excitation vocoders, exploit the phase redundancy in the HAS. The HAS is relatively sensitive to the distortion of spectral magnitudes. However, for the distortion of the phase, the HAS exhibits partial deafness. Thus, most of sinusoidal speech coders transmit just the spectral magnitudes of the residual signal, and generate the phase components using the pitch and the voicing levels for each harmonic in synthesising speech signals at the decoder. For voiced harmonics, the phase can be reconstructed by an interpolation method using the initial phase with the pitches of the last and current frames [66] [37]. However, random phases are used for unvoiced harmonics.

Based on the redundancy being exploited, speech coders are classified as given in Table 5.1, which shows that fundamental progress in speech coding techniques is mainly due to successful introduction of statistical and psychoacoustic redundancies in human speech. Additionally, the advances in elaborate estimation and quantisation techniques for the model parameters contribute to attaining consistent speech-quality.

Recently, many low bit-rate speech coders below 4 kbit/s are investigated based on the sinusoidal model [69][70][71][95][16]. Thus, it is worthwhile to discuss more on the detailed structure of sinusoidal speech coding techniques as in the following subsection.
Table 5.1: Classification of speech coders according to redundancy features being exploited.

<table>
<thead>
<tr>
<th>Psychoacoustic redundancy</th>
<th>Statistical redundancy</th>
<th>STC</th>
<th>STC + LTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>PCM (64 kbit/s)</td>
<td>ADPCM (32 ~ 16 kbit/s)</td>
<td>-</td>
</tr>
<tr>
<td>PW</td>
<td>-</td>
<td>-</td>
<td>CELP, RPE-LPC, MPE-LTP (16 ~ 4 kbit/s)</td>
</tr>
<tr>
<td>PW + phase</td>
<td>-</td>
<td>-</td>
<td>Sinusoidal coding (~ 4 kbit/s)</td>
</tr>
</tbody>
</table>

STC: short-term correlation, LTC: long-term correlation, PW: perceptual weighting

5.4 Sinusoidal Model-based Speech Coding

5.4.1 Speech Modelling

Speech signals can be synthesised by the sum of sinusoidal signals of which the frequencies are sampled at harmonic frequencies of the spectral envelope as

\[ s(n) = g \sum_{h=1}^{H} A_{ip}(h\omega_0) \cos(h\omega_0 n + \theta_h + \theta_0) \]  

(5.3)

where \( g, H, \omega_0, \theta_0, \) and \( A_{ip}(-) \) are the gain, the total number of harmonics, the fundamental frequency, the initial phase, and the spectral envelope magnitudes, respectively. The phase \( \theta_h \) for the \( h \)th harmonic component is

\[ \theta_h = \begin{cases} 0, & \text{if } v_h \text{ is voiced} \\ U(-\pi, \pi), & \text{otherwise} \end{cases} \]  

(5.4)

in which \( v_h \) is the voicing state, either voiced or unvoiced, and \( U[\theta_1, \theta_2] \) is the uniform random number with distribution in the range of \([\theta_1, \theta_2] \).

However, in recent sinusoidal coders, the sinusoidal signals are synthesised for excitation signals rather than speech signals [72] [36] as

\[ e(n) = \sum_{h=1}^{H} A_e(h\omega_0) \cos(h\omega_0 n + \theta_h + \theta_0) \]  

(5.5)
where $A_e(\cdot)$ is the spectral magnitude of the residual signal. The motivation behind this approach is that the residual signal compared to the speech signal is more suitable for modelling with zero and random phases. The speech signal can be synthesised by the convolution between $\hat{e}(n)$ in (5.5) and the LP synthesis filter $h(n)$, as

$$s(n) = \hat{e}(n) * h(n).$$  \hspace{1cm} (5.6)

Thus, the model parameters required to estimate, quantise, and transmit in sinusoidal speech coding are typically composed of

- LP coefficients, $\{a_i\}_{i=0}^{p}$
- Pitch, $T_0$ (or $\omega_0$)
- Voicing decision, $v_h$ for $h = 1, \cdots, H$
- Spectral magnitudes, $A_e(h\omega_0)$ for $h = 1, \cdots, H$
- Gain, $g$

Note that LPC vocoders can be considered as a special case of sinusoidal speech coders, if the voicing decision is conducted over the full-band rather than more than two sub-bands, and the magnitude spectrum is flat. However, the sinusoidal coder becomes a CELP-like coder if the spectral phases are included in the model parameter. Thus, sinusoidal coders can be classified as an intermediate approach between LPC and CELP coders, aiming to achieve toll quality speech at a low bit-rate by exploiting the partial phase deafness of the HAS.

### 5.4.2 LP-coefficient Estimation

LP-coefficient $a_t$ is obtained from the autocorrelation coefficients followed by Levinson-Durbin recursion \cite{73} for $m = 1, 2, \cdots, p$, as

$$k_m = \frac{R(m) - \sum_{k=1}^{m-1} a_{m-1}(k) R(m - k)}{E_{m-1}},$$

$$a_m(m) = k_m,$$
5.4. Sinusoidal Model-based Speech Coding

\[ a_m(k) = a_{m-1}(k) - k_m a_{m-1}(m-k), 1 \leq k \leq m-1, \]

\[ E_m = (1 - k_m^2)E_{m-1}, \tag{5.7} \]

where \( E_0 = R(0) \) and \( R(m) \) and \( k_m \) are the autocorrelation and reflection coefficients, respectively. Now, the LP synthesis filter \( H(z) \) is given

\[ H(z) = \frac{1}{A_p(z)} = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \tag{5.8} \]

where \( a_0 = 1 \). The LP coefficients are transformed to the line spectral pair (LSP) [74] for effective quantisation and interpolation [75]. The LSP coefficients are defined as the roots of the polynomials,

\[ P(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \tag{5.9} \]

and

\[ Q(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \tag{5.10} \]

where \( A_p(z) \) represents the LP inverse filter with the order \( p \). If \( p \) is even, typically used in speech coders, \( P(z) \) and \( Q(z) \) can be factorised as

\[ P(z) = (1 - z^{-1}) \prod_{i=1,2,\ldots,p/2} (1 - 2 p_i z^{-1} + z^{-2}) \tag{5.11} \]

and

\[ Q(z) = (1 + z^{-1}) \prod_{i=1,2,\ldots,p/2} (1 - 2 q_i z^{-1} + z^{-2}) \tag{5.12} \]

where \( p_i = \cos \omega_{2i} \) and \( q_i = \cos \omega_{2i-1} \), \( 1 \leq i \leq p/2 \), are referred to as the LSPs of the line spectral frequencies (LSF) \( \omega_j \), \( 1 \leq j \leq p \), which are ordered as \( 0 < \omega_1 < \omega_2 < \cdots < \omega_p \leq \pi \). The roots of \( P(z)/(1 - z^{-1}) \) and \( Q(z)/(1 + z^{-1}) \) can be obtained by Newton's iteration method after manipulating the polynomials in an appropriate manner. The transformation of LSP to LP-coefficient can be carried out by comparing the impulse responses of the LSP and LP-synthesis filters.
5.4.3 Pitch Estimation

The pitch is one of the most important parameters in sinusoidal speech coders. Incorrect estimation of the pitch leads to wrong estimates of the other model parameters, since most of the parameters can be estimated with a correct pitch.

In the multiband excitation (MBE) vocoder [67], the pitch $\omega_0$ is estimated using a spectral synthesis method in terms of minimising the error metric $e^{(MBE)}(\tau)$ as

$$e^{(MBE)}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_w(\omega) - \hat{S}_w(\omega)|^2 d\omega$$  \hspace{1cm} (5.13)

where $S_w(\omega)$ is the spectrum of windowed speech signals; and $\hat{S}_w(\omega)$ is the synthesised spectrum given

$$\hat{S}_w(\omega) = \sum_{h=1}^{H(\tau)} H_w(\omega)W(\omega - h\omega_0)$$  \hspace{1cm} (5.14)

where $\tau = 2\pi/\omega$ and $H(x) = |x/2|$. In (5.14), $H_w(\cdot)$ and $W(\cdot)$ are the spectral envelope and the applied window spectrum, respectively.

The sinusoidal transform coder (STC) [66] also conducts pitch estimation using the spectral synthesis technique by maximising the error metric $e^{(STC)}(\tau)$, as

$$e^{(STC)}(\tau) = \sum_{h=1}^{H(\tau)} A\left(\frac{hK}{\tau}\right) \left\{ \sum_{k=1}^{K/2} A(k)D\left(\frac{k}{K} - \frac{h}{\tau}\right) - \frac{1}{2} A\left(\frac{hK}{\tau}\right) \right\}$$  \hspace{1cm} (5.15)

where $D(x) = \frac{\sin(2\pi x)}{2\pi x}$.

5.4.4 Voicing Estimation

Frequency-dependent voicing estimation is a unique feature of sinusoidal speech coders. There are various ways how to define the voicing and to estimate its level. In sinusoidal speech coding, the excitation signal is represented by the mixture of voiced and unvoiced signals, rather than a switch logic between the two excitations.

The concept for mixed excitation between periodic and random signals was first proposed by Makhoul [76], who proposed periodic excitation for lower frequencies and random excitation for higher frequencies in which the frequency separating the two
bands is defined by a cutoff. He discussed the method, however, in conceptual level without a proper experiment.

Kwon extended the LPC vocoder with the mixture control scheme [77] as

\[ e(n) = \alpha e_v(n) + (1 - \alpha)e_u(n) \]  (5.16)

where \( e_v(t) \) and \( e_u(t) \) mean the voiced and unvoiced excitations, respectively, and \( \alpha \) ranging between 0 and 1 is the rate for the mixture control. LPC vocoders are a special case of the Kwon's speech coder, if \( \alpha \) has just a binary value of either 1 or 0. The sound becomes strongly voiced if \( \alpha \) is close to 1. On the other hand, by decreasing \( \alpha \), it is possible to produce white noise-like excitation.

Griffin in the MBE vocoder proposed a voicing scheme depending on frequencies [67] as

\[ e(n) = \sum_{h=1}^{H} v_h e_v(n) + (1 - v_h)e_u(n) \]  (5.17)

where

\[ v_h = \begin{cases} 
1 & \text{if } h\text{th harmonic is voiced,} \\
0 & \text{otherwise.}
\end{cases} \]

However, the number of bands \( H \) are too large to transmit the voicing levels for each harmonic. Thus, the MBE vocoder quantises voicing information by clustering the spectral harmonics into a small number of sub-bands.

It is a well-known fact that speech signals exhibit voiced harmonic for relatively lower harmonics and unvoiced for higher ones [78] [79]. Thus, recent sinusoidal speech coders estimate a cutoff frequency \( H_c \) separating the voiced and unvoiced harmonic components for the lower and higher frequencies, respectively, as

\[ e(t) = \sum_{k=1}^{H_c} e_v(t) + \sum_{k=H_c+1}^{H} e_u(t). \]  (5.18)

The cutoff harmonic frequency scheme is widely adopted by many sinusoidal speech coders, because it is possible to allocate a smaller number of bits without speech quality degradation [80] [79] [81].
5.4. Sinusoidal Model-based Speech Coding

STC estimates the voicing cutoff based on the signal fit between the original and synthesised speech signals [80], as

$$SNR = \frac{\sum |s(n)|^2}{\sum |s(n) - \tilde{s}(n; T_0)|^2}$$

(5.19)

where $\tilde{s}(n; T_0)$ is the synthesised speech signal for the estimated pitch $T_0$. Through proper derivation steps, it can be rewritten as

$$SNR = \frac{P_s}{P_s - 2c^{(STC)}(T_0)}$$

(5.20)

where $c^{(STC)}(T_0)$ is defined in (5.15), and $P_s$ is the power sum of each harmonic amplitude. The voicing cutoff is defined with $SNR$ as

$$P_v(SNR) = \begin{cases} 1, & \text{if } SNR > 13 \text{ dB}, \\ \frac{1}{2}(SNR - 4), & \text{if } 4 \text{ dB} \leq SNR \leq 13 \text{ dB}, \\ 0, & \text{if } SNR < 4 \text{ dB}. \end{cases}$$

(5.21)

Finally, the harmonic index for the voicing cutoff is given

$$H_c = \lfloor P_v(SNR)H \rfloor.$$  

(5.22)

5.4.5 LP-coefficient Quantisation

Quantisation of the LP-coefficient is a traditional task in low bit-rate speech coders in order to transmit the coefficients using a small number of bits while keeping a good speech-quality. The LSF instead of the LP-coefficient is widely used in quantisation due to its ordering property [75]. It is possible to design an efficient quantiser as the distribution of each LSF is localised mostly to a certain range. Furthermore, the ordered sequence of the LSFs guarantees the stability of the filter.

Scalar quantisation can be conducted by independently quantising each LSF. However, it is possible to attain better rate-distortion by employing a vector quantiser according to the information theory. But, it is practically not feasible to design a universal LSF vector quantiser (VQ), covering the whole vector space using a codebook. Thus various tractable algorithms, such as split vector quantisation (SVQ) [82] and multi-stage vector quantisation (MSVQ) [83], are proposed.
SVQ [82] decomposes the vector space into a certain number of subvectors, and then quantises each subvector independently. Given a pth order LSF vector, \( \mathcal{J} = (\omega_1, \omega_1, \ldots, \omega_p)^T \), SVQ decomposes \( \mathcal{J} \) into \( L (< p) \) subvectors as

\[
\mathcal{J}^{(SVQ)} = (\bar{\omega}_1, \bar{\omega}_2, \ldots, \bar{\omega}_L)
\]

where \( \bar{\omega}_1 = (\omega_1, \ldots, \omega_{u_1})^T \), \( \bar{\omega}_2 = (\omega_{u_1+1}, \ldots, \omega_{u_2})^T \), \ldots, \( \bar{\omega}_L = (\omega_{u_{L-1}+1}, \ldots, \omega_p)^T \) in which \( u_j \) means the topmost index of the \( j \)th subvector. SVQ independently quantises each \( \bar{\omega}_i \), \( 1 \leq i \leq L \), using a typical VQ. In order to exploit the correlation between the subvectors, Kim, et al. proposed linked SVQ (LSVQ) [84], which builds more than two codebooks for each subvector except the initial subvector. Let \( \bar{\omega}_k \) be the initial subvector. LSVQ selects proper codebooks for \( k \)-1th and \( k \)th subvectors depending on the values of \( \omega_{u_{k-1}+1} \) and \( \omega_{u_k} \), respectively. In the same way, it searches optimum codebooks for each remaining subvector. LSVQ exploits the ordering property of LSFs, and presents lower spectral distortion compared with SVQ [84].

MSVQ [83] conducts LSF quantisation using a set of cascaded VQs composing multiple stages. The final quantised LSF \( \mathcal{J} \) is the sum of the intermediate outputs of each VQ stage as

\[
\hat{\mathcal{J}} = \sum_{i=1}^{M} \tilde{\mathcal{J}}^{(i)}
\]

where \( M \) is the number of the total stages, and \( \tilde{\mathcal{J}}^{(i)} \) is the intermediate result of the \( i \)th quantisation stage. The vector of each stage is quantised by a pth order VQ. For the \( j \)th stage, the residual vector being quantised is \( \mathcal{J} - \sum_{i=1}^{j-1} \tilde{\mathcal{J}}^{(i)} \).

In consideration of time and space complexities, SVQ and MSVQ present tractable solutions for quantisation of LSF vectors. It is also possible to design a VQ with the combination of SVQ and MSVQ [12]. At the first stage, it quantises the input vector using a universal vector quantiser, and then at the second stage it quantises the residual vector using SVQ because SVQ is useful for the vectors with a lower intraframe correlation compared to MSVQ.

Predictive coding schemes including moving average (MA), auto-regressive (AR), and switched-predictive (SP), have been investigated intensively for the first stage of MSVQ in order to exploit interframe redundancy in LSF vectors [85] [86] [87].
5.4. Sinusoidal Model-based Speech Coding

The LSF residue $\Delta \omega^{(MA)}(t)$ by first-order MA-based prediction [85] is given by

$$\Delta \omega^{(MA)}(t) = \omega(t) - P^{(MA)} \Delta \omega^{(MA)}(t - 1)$$

(5.25)

in which $t$ is the frame index, and $P^{(MA)}$ is the $p \times p$ MA-prediction matrix. Subsequently, the residual vector can be quantised using a vector quantiser $Q[\cdot]$, e.g. MSVQ, as

$$\Delta \tilde{\omega}^{(MA)}(t) = Q [\Delta \omega^{(MA)}(t)]$$

(5.26)

Finally, the LSF can be reconstructed as

$$\tilde{\omega}^{(MA)}(t) = \Delta \tilde{\omega}^{(MA)}(t) + P^{(MA)} \Delta \omega^{(MA)}(t - 1).$$

(5.27)

The MA-based scheme guarantees stability due to the nature of the filter. Furthermore, under bit-error prone channel environments, the propagation of the error is limited to the order of MA-prediction.

Using AR-based prediction, it is also possible to quantise the LSF as

$$\Delta \tilde{\omega}^{(AR)}(t) = Q [\tilde{\omega}(t) - P^{(AR)} \tilde{\omega}^{(AR)}(t - 1)]$$

(5.28)

in which $P^{(AR)}$ is the $p \times p$ AR-prediction matrix. The LSF can be reconstructed as

$$\tilde{\omega}^{(AR)}(t) = \Delta \tilde{\omega}^{(AR)}(t) + P^{(AR)} \tilde{\omega}^{(AR)}(t - 1).$$

(5.29)

The AR-based scheme is effective for removing interframe redundancy compared with the MA-based scheme. However, it cannot be applied to error-prone channel environments, because the error affects the subsequent frames in reconstruction of the LSF.

To mitigate the channel error problem while achieving a relatively low spectral distortion, the switched predictive (SP) scheme [86] [87] is introduced as

$$\Delta \tilde{\omega}^{(SP)}(t) = \begin{cases} 
Q [\tilde{\omega}(t) - \tilde{\omega}(t)] & \text{if } M(t) = AR \\
Q [\tilde{\omega}(t)] & \text{if } M(t) = ML
\end{cases}$$

(5.30)

where $\tilde{\omega}(t)$ is the predicted term by the previously reconstructed LSF $\tilde{\omega}^{(SP)}(t - 1)$ with the AR-predictor $P^{(AR)}$ as

$$\tilde{\omega}(t) = P^{(AR)} \tilde{\omega}^{(SP)}(t - 1).$$

(5.31)
In (5.30), $M(\cdot)$ indicates the switch mode, defined
\[
M(t) = \begin{cases} 
\text{AR} & \text{if } d(\hat{\omega}(t), \hat{\omega}(t)) < d(\hat{\omega}(t), Q[\hat{\omega}(t)]) \\
\text{ML} & \text{otherwise}
\end{cases}
\] (5.32)
where $d(x, y)$ is a distortion measure between $x$ and $y$. The SP scheme quantises the predicted residual using the memory information in the case of the AR mode which is useful for stationary regions of speech. For transitional regions, LSF exhibits high fluctuation from frame to frame. In this case, it is possible to attain lower spectral distortion by switching it to the ML mode. Additionally, a reset scheme enforcing to select the ML mode regardless of a measured distortion can be applied every a certain number of frames in order to further restrict the AR-based memory effect. The LSF can be reconstructed finally as
\[
\hat{\omega}^{(\text{SP})}(t) = \begin{cases} 
\Delta \hat{\omega}^{(\text{SP})}(t) + \hat{\omega}(t) & \text{if } M(t) = \text{AR} \\
\Delta \hat{\omega}^{(\text{SP})}(t) & \text{if } M(t) = \text{ML}.
\end{cases}
\] (5.33)

5.4.6 Spectral Magnitudes Quantisation

The spectral magnitudes of the LP residual signal is crucial to enhance the speech quality of sinusoidal coders since the LP-spectral envelope with a low model-order could not flatten the shape of spectral magnitudes. Thus, various methods for efficient quantisation of the spectral magnitudes have been investigated by many researchers. The traditional problems in this task is the vector dimension which varies, from frame to frame, depending on the pitch.

The improved MBE vocoder [37] transforms the magnitude spectrum using the discrete cosine transform, and then quantises the coefficients with the combination of scalar and vector quantisers [37]. The sinusoidal transform coder represents the spectrum with a high order all-pole model [80]. In band-limited interpolation [88][78], the variable-dimension of the spectrum is converted into a fixed-dimension based on the sampling rate conversion and signal interpolation techniques. In variable dimension vector quantisation [89], the spectral vector is quantised directly using a universal codebook of a fixed-dimension, in which each element of the spectral vector is mapped onto a code vector using a selector. In non-squared transform vector quantisation [90], the input
vector is transformed into a fixed-dimension using a linear transform matrix. Additionally, the quantisation performance can be improved perceptually by incorporating the perceptual weighting scheme in the distortion measure [70][91].

5.5 Pitch Estimation using Spectral Autocorrelation

5.5.1 Introduction

The human voice signals are formed by passing quasi-periodic excitation signals through the vocal filter. The duration between the excitation signal is called the pitch period $T_0$ or fundamental frequency $f_0$. Estimation of the pitch period is essential in most of low bit-rate speech coders in order to exploit the quasi-periodic redundancy in human speech. However, incorrect estimation of the pitch period could seriously degrade the quality of the synthesised speech.

Pitch determination algorithms (PDAs) have been studied mostly in the time or frequency domain. Comparison among the methods is discussed in [92][93]. Traditionally, autocorrelation-based methods [94] with its variants [95][96] have been intensively investigated and widely applied to various speech coders [10][12][20][21][22][35]. The frequency domain approaches [97][67][80] have been fueled recently by growing interests in sinusoidal speech coders, such as the multi-band excitation (MBE) [67] and the sinusoidal transform coder (STC) [80], which conduct pitch determination based on a spectral synthesis (SS) method.

Pitch estimation errors can be classified into fine and gross pitch errors [93]. The fine pitch error means that the pitch precision is not high enough that the human auditory system can perceive the difference of the pitch resolution. The coarse pitch can be refined using fractional pitch estimation techniques in the time or frequency domain [98][99][37]. The gross pitch error, including pitch period multiples and submultiples, means that the difference between the true and the estimated pitch periods is relatively large, e.g. pitch doubling and halving.

In order to tackle the gross pitch errors, pitch tracking methods can be applied, in consideration of the fact that the human pitches change slowly from frame to frame. In
other words, it is possible to determine a smoothed version of the current frame's pitch using the past and future pitch candidates by means of a tracking method, such as the median filter [100] or the dynamic programming technique [67]. The pitch tracking method, however, can be limited or cannot be feasible depending on application areas as it requires pitch candidates in the future speech frames introducing additional delay. Thus, it cannot be deployed in telecommunications systems requiring a short delay for natural two-way communications. In most of voice communications, the problems in pitch estimation need to be alleviated without introducing any extra delay.

The autocorrelation-based PDA would give rise to high correlation values over integer multiple lags of the pitch period. A well known technique for alleviating the pitch period multiple problem is the weighting method [101], putting preference to smaller lags of autocorrelation, which is adopted by many standard speech coders [12][21][22]. It is, however, not easy to find an optimum weight over various kinds of speech signals, which becomes even more difficult in noise environments.

The pitch period submultiples are mainly related with the effect of the formants of speech, especially when the first formant \( F_1 \) coincides with multiples of \( f_0 \). Pitch estimation metrics would mis-detect \( F_1 \) as \( f_0 \).

In this section, we propose a novel pitch estimation metric based on spectral autocorrelation (SA), exploiting the autocorrelation of harmonic magnitude spectrum, so as to attain considerable reduction in the number of gross pitch errors.

This section is organised as follows. Subsection 5.5.2 defines SA and describes its characteristics. Next, in subsection 5.5.3, two kinds of PDAs incorporating SA are proposed for alleviating the pitch multiple problems. In subsection 5.5.3.3, a spectral flattening technique for suppressing the formant effect is discussed. Finally, subsection 5.5.4 shows experimental results, detailing performance evaluation of the proposed PDAs with optimisation of weighting factors being used.

### 5.5.2 Spectral Autocorrelation

Excitation signals for voiced speech exhibit harmonics in the magnitude spectrum as the voiced speech is generated by a time-varying filter excited by a quasi-periodic impulse-
5.5. Pitch Estimation using Spectral Autocorrelation

The distance between the two consecutive harmonics is the fundamental frequency \( f_0 \). Autocorrelation conducted over the magnitude spectrum, called spectral autocorrelation (SA), is considered for the measurement of the pitch period.

Let a spectrum \( S(m) = A(m)e^{j\theta(m)} \) for \( 0 \leq m \leq M - 1 \), where \( A(m) \) and \( \theta(m) \) are the magnitude and phase spectra for a segment of windowed signal, and \( M \) is the number of discrete Fourier transform (DFT) points. Normalized SA \( R_\Sigma(\tau) \) is defined as

\[
R_\Sigma(\tau) = \frac{\sum_{m=0}^{[M/2]} A_x(m)A_x(m+\omega_\tau)}{\sqrt{\left\{ \sum_{m=0}^{[M/2]} A_x^2(m) \right\} \left\{ \sum_{m=0}^{[M/2]} A_x^2(m+\omega_\tau) \right\}} , \text{ for } T_0^{(l)} \leq \tau \leq T_0^{(u)} \tag{5.34}
\]

where \( \omega_\tau = \left[ \frac{M}{\tau} + 0.5 \right] \); and \( T_0^{(l)} \) and \( T_0^{(u)} \) are the lower and upper numbers of samples for the pitch search. In (5.34), the zero-crossing spectrum \( A_x(m) \) is given by

\[
A_x(m) = A(m) - g\bar{A}(m) \tag{5.35}
\]

where \( \bar{A}(m) \) is the spectral envelope of \( A(m) \). The envelope is estimated using the peak-picking method [71][102]. The magnitude spectrum \( A(m) \) is converted into the zero-crossing spectrum \( A_x(m) \) to make it feasible for the autocorrelation defined in (5.34). The gain \( g \) is calculated as \( g = \sum_{m=0}^{[M/2]} A(m)\bar{A}(m)) / \sum_{m=0}^{[M/2]} A(m)A(m) \). In (5.35), the logarithmic spectrum could be considered to obtain a zero-crossing spectrum. However, the SA with the logarithmic spectrum produces high correlation ratio for large lags \( \tau \) close to \( T_0^{(u)} \) corresponding to very small \( \omega_\tau \)-s, i.e. \( [M/T_0^{(u)} + 0.5] \leq \omega_\tau \ll [M/(2T_0) + 0.5] \). Thus, the linear magnitude spectrum is used instead of the logarithmic one.

Fig. 5.1 shows an example illustrating the characteristics of the SA. For a speech segment in Fig. 5.1(a), the magnitude and its zero-crossing spectra are shown in Fig. 5.1(b) and (c), respectively. Finally, spectral autocorrelation is shown in Fig. 5.1(d), indicating a prominent peak at the pitch lag.

The autocorrelation over a harmonic spectrum produces high correlation for integer multiple lags in the frequency domain. However, the corresponding time-domain lag of the SA exhibits the following unwanted property,
Figure 5.1: An example of (a) speech signal of $T_0 = 34$-sample ($F_s = 8$ kHz), (b) magnitude spectrum, (c) zero-crossing spectrum, and (d) spectral autocorrelation.
Property 1 (Submultiplicity): For a segment of periodic impulse-train-like signal with the fundamental period $T_0 \geq T_0^{(l)}$, $R_S(\tau)$ in (5.34) has peaks for the integer submultiples of $T_0$, i.e. $\tau = T_0/k$, for $1 \leq k \leq \lfloor T_0/T_0^{(l)} \rfloor$.

Autocorrelation applied directly to the harmonic spectrum features relatively high values at the integer multiple lags $kf_0$, for $1 \leq k \leq \lfloor F_s/F_0 \rfloor$, where $F_s$ is the sampling frequency. Each $kf_0$ in the frequency scale corresponds to $T_0/k$ in the time scale. Thus, $R_s(\tau)$ features strong peaks for the integer submultiples of $T_0$.

Fig. 5.2 shows an example featuring high SAs for pitch period submultiples. The SA can be used to detect the pitch especially for high-pitched speech signals, but it exhibits the drawback for low-pitched signals. Even worse, the large lags of SA exhibit low pitch-precision depending on the number of DFT points. Thus, the SA metric may not be used as an independent PDA, but could be useful when combined with other metrics because it does not give rise to the pitch period multiple problem.

5.5.3 Pitch Estimation using Spectral Autocorrelation

5.5.3.1 A PDA based on Spectro-Temporal Autocorrelation

Autocorrelation conducted over time-domain signals is named temporal autocorrelation (TA). Given a segment of speech signals $s(n)$, $0 \leq n \leq N-1$, the normalised TA for a pitch candidate $\tau$ in the number of waveform samples is given by

$$R_T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} s(n)s(n+\tau)}{\sqrt{\left\{ \sum_{n=0}^{N-\tau-1} s^2(n) \right\} \left\{ \sum_{n=0}^{N-\tau-1} s^2(n+\tau) \right\}}}.$$  \hspace{1cm} (5.36)

The TA has been widely used for PDAs due to its relatively good performance especially over noisy speech signals [93]. However, it has an undesirable property in terms of reliable pitch estimation as below,

Property 2 (Multiplicity): For a segment of periodic signals with a fundamental period $T_0 \leq T_0^{(u)}$, $R_T(\tau)$ in (5.36) has peaks for the integer multiples of $T_0$, i.e. $\tau = kT_0$ for $1 \leq k \leq \lfloor T_0^{(u)}/T_0 \rfloor$. 

Figure 5.2: An example of (a) speech signal of $T_0 = 59$-sample ($F_s = 8$ kHz), (b) magnitude spectrum, (c) zero-crossing spectrum, and (d) spectral autocorrelation.
Periodic signals with a period $T_0$ are also periodic for integer multiple periods of $T_0$. Thus, $R_T(\tau)$ of the periodic signal becomes large for the integer multiple lags of $T_0$.

Thus, the TA-based PDA would result in detecting an unwanted pitch period multiple, according to Property 2. Since the number of pitch period multiples increases with decreasing $T_0$ for a given search interval, higher-pitched speech signals have higher potential for resulting in pitch estimation error.

However, the TA is useful for detecting the pitch of low-pitched speech while the SA for high-pitched speech. In other words, the pitch period multiple and submultiple problems can be compensated if the two autocorrelation methods, TA and SA, are combined in an advantage way. Hence, spectro-temporal autocorrelation (STA) is defined as

$$R_{ST}(\tau) = \alpha R_T(\tau) + (1 - \alpha) R_S(\tau)$$

(5.37)

where $\alpha$ is the weighting factor, $0 \leq \alpha \leq 1$. The cases of $\alpha = 0$ and $\alpha = 1$ reduce the STA to the SA and the TA, respectively. The estimated pitch period $\hat{T}_0$ using the STA is the argument maximising (5.37) as

$$\hat{T}_0 = \arg \max_{\tau} \{ R_{ST}(\tau) \}. \quad (5.38)$$

Because of the dual relation between the temporal and the spectral autocorrelations, it is found that the STA has a useful property for pitch estimation,

**Property 3 (Duality) :** For a segment of periodic impulse-train-like signals with a fundamental period $T_0$, $T_0^{(l)} \leq T_0 \leq T_0^{(u)}$, $R_{ST}(\tau)$ in (5.37) has a strongest peak at $\tau = T_0$ compared with the integer multiple and submultiple periods of $T_0$, i.e. $\tau = pT_0$ and $\tau = T_0/q$ for $2 \leq p \leq \lfloor T_0^{(u)}/T_0 \rfloor$ and $2 \leq q \leq \lfloor T_0^{(l)}/T_0 \rfloor$.

As discussed, $R_S(\tau)$ in (5.34) and $R_T(\tau)$ in (5.36) do not exhibit high peaks for the integer multiples and submultiples of $T_0$, i.e. $\tau = kT_0$ and $\tau = T_0/k$ for $k \geq 2$, respectively, for periodic impulse-train-like signals. In (5.37), $R_S(\tau)$ and $R_T(\tau)$ terms suppress the undesirable high peaks for the multiples and submultiples of $T_0$ excluding
Table 5.2: Comparison of the number of strong peaks of periodic impulse-train-like signals.

<table>
<thead>
<tr>
<th>Pitch range (samples)</th>
<th>TA</th>
<th>SA</th>
<th>STA</th>
</tr>
</thead>
<tbody>
<tr>
<td>High: ([T_0^{(l)}, 2T_0^{(l)} - 1])</td>
<td>(\geq 2)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mid: ([2T_0^{(l)}, T_0^{(u)}/2])</td>
<td>(\geq 2)</td>
<td>(\geq 2)</td>
<td>1</td>
</tr>
<tr>
<td>Low: ([T_0^{(u)}/2 + 1, T_0^{(u)}])</td>
<td>1</td>
<td>(\geq 2)</td>
<td>1</td>
</tr>
</tbody>
</table>

\(\tau = T_0\), i.e. \(\tau = kT_0\) and \(\tau = T_0/k\) for \(k \geq 2\), respectively. Consequently, the STA for \(\tau = T_0\) remains relatively more prominent compared with those for the rest.

The range of the pitch period can be split into three groups as high, mid, and low-pitched, based on the expected number of prominent peaks in the TA and SA. The minimum pitch period producing a pitch period submultiple in SA is \(2T_0^{(l)}\). The SA is capable of rarely producing pitch period submultiples for high-pitched signals, i.e \(T_0^{(l)} \leq T_0 \leq 2T_0^{(l)} - 1\). On the other hand, in the TA, the maximum pitch period generating the pitch period multiple is \(T_0^{(u)}/2\). Thus, for \(T_0^{(u)}/2 + 1 \leq T_0 \leq T_0^{(u)}\), the TA can be relatively robust against the gross pitch error. However, the STA gives robust results for the whole pitch range by combining the two functions. Comparison of the number of expected peaks among the TA, SA, and STA for each pitch range is shown in Table 5.2 in which the effect due to the formant resonance is not considered but is discussed in Section 5.5.3.3.

Examples comparing the characteristics of the TA, SA, and STA for speech signals with various pitch periods are shown in Fig. 5.3. Each TA and SA exhibits a unique strong peak, in Fig. 5.3(f) and (g), for low and high-pitched speech signals, in Fig. 5.3(c) and (a), respectively. However, both TA and SA feature high values at multiple lags, as shown in Fig. 5.3(d) and (i), for the high and low-pitched speech signals, respectively. For the mid-pitched speech signal in Fig. 5.3(h), both TA and SA exhibit multiple strong peaks in Fig. 5.3(e) and (h), respectively. However, the STA exhibits a robust peak regardless of the pitch period, as shown in Fig. 5.3(j), (k), and (l).
Figure 5.3: Comparison of TA, SA, and STA ($\alpha = 0.5$) for (a) high, (b) mid, and (c) low-pitched speech signals with pitch periods 32, 59, and 100 samples, respectively.
5.5. Pitch Estimation using Spectral Autocorrelation

5.5.3.2 A PDA based on Spectral Synthesis incorporating Spectral Autocorrelation

The spectrum for windowed speech signals can be decomposed into the spectral envelope and the excitation spectrum. The spectral envelope is the smoothed version of the speech spectrum. The excitation spectrum exhibits harmonics for voiced components, in which each harmonic typically has the shape of the sinc function corresponding to the frequency response of the applied window. Spectral synthesis (SS) methods [67][80] find the pitch so as to minimise the distortion between the original and synthesised spectra. The synthesised spectrum is generated by shifting the centre frequency of the sync function spectrum to harmonic frequencies.

In [80], the McAulay's SS-based PDA, the metric for pitch determination is given by

\[
\Gamma' (\tau) = \sum_{h=1}^{H} h K A' (\frac{h K}{\tau}) \left\{ \sum_{k=1}^{K/2} A(k) D \left( \frac{h}{K} - \frac{h}{\tau} \right) - \frac{1}{2} A(hK) \right\} \tag{5.39}
\]

where \( H(x) = \lfloor x/2 \rfloor \) and \( D(x) = \sin(2\pi x)/(2\pi x) \). It can be improved against the formant effect by incorporating an energy-based metric \( \varphi(\tau) \) [71][81], given by

\[
\varphi(\tau) = \sum_{n=0}^{N} |d_{\tau}(n)| \sum_{n=0}^{N} |e_{\tau}(n)| \tag{5.40}
\]

where \( e_{\tau}(n) = \sum_{k=0}^{N} s_{\tau}(n - \lfloor \tau/2 \rfloor + k) \) and \( d_{\tau}(n) = 0.95d_{\tau}(n - 1) + e_{\tau}(n) - e_{\tau}(n - 1) \) with \( d_{\tau}(0) = 0 \). Finally, the SS-based metric is defined as

\[
\Gamma_{\varphi}(\tau) = \frac{\Gamma'(\tau)}{\varphi(\tau)} \tag{5.41}
\]

in which \( \Gamma_{\varphi}(\tau) \), if not positive, is bounded to a small positive value.

The SS-based method can be further improved by incorporating the spectral autocorrelation metric in (5.34) since \( \Gamma_{\varphi}(\tau) \) could also exhibit high peaks at pitch period multiples like the TA. Hence, SS incorporating SA, called SS-SA, is defined as

\[
\Gamma_{SA}(\tau) = \left\{ \Gamma_{\varphi}(\tau) \right\}^\beta \left( \frac{R_{S}(\tau) + 1}{2} \right)^{1-\beta} \tag{5.42}
\]

where \( \beta \) is the weighting factor, \( 0 \leq \beta \leq 1 \). The SS-SA becomes SS and SA for the cases of \( \beta = 1 \) and \( \beta = 0 \), respectively.
5.5. Pitch Estimation using Spectral Autocorrelation

Figure 5.4: Comparison between SS and SS-SA ($\beta = 0.25$) for high, mid, and low-pitched signals shown in Fig. 5.3(a), (b), and (c), respectively. The SAs are shown in Fig. 5.3(g), (h), and (i).

Examples, examining the characteristics of $\Gamma_\varphi(\tau)$ and $\Gamma_{ST}(\tau)$, are shown in Fig. 5.4 in which the measured value of each sub-figure is normalised by each maximum value. The input speech signals used in Fig. 5.4 are same to the ones used in the analysis of the STA. For the high-pitched signals in Fig. 5.3(a), the lag corresponding to the pitch period double exhibits a strong peak in Fig. 5.4(a), even stronger than the peak of the correct pitch. The SS-SA alleviates this problem as shown in Fig. 5.4(d). The peak of the pitch lag becomes prominent in comparison with those of other lags. For the mid and low-pitched signals in Fig. 5.3(b) and (c), the maximum peaks of $\Gamma_\varphi(\tau)$ and $\Gamma_{ST}(\tau)$ are relatively obvious as illustrated in Fig. 5.4(b), (c), (e), and (f).

5.5.3.3 Spectral Flattening

Formant interaction with the fundamental frequency can cause problems in pitch estimation. Several methods have been proposed to flatten the speech spectrum in order
to avoid the formant interaction effect [94][96][103][104][81]. One popular method is the use of linear predictive (LP)-inverse filtering [96][104]. However, this technique may over-estimate the harmonics structure, particularly in high-pitched speech signals. Thus, a formant weighting filter [60] is adopted as a preprocessor to control the de-emphasising factor of the formants while keeping the harmonics structure. It has an advantage of taking the intermediate signal between the speech and LP-residual signals as

\[
S_f(z) = \frac{A(z)}{A(z/\gamma)} S(z)
\]

(5.43)

where \( S(z) \), \( S_f(z) \), and \( A(z) \) are the \( z \)-transform of the input speech signal \( s(n) \), the formant suppressed signal, and the inverse filter, respectively. The parameter \( \gamma \) is the formant weighting factor, \( 0 \leq \gamma \leq 1 \). For the case of \( \gamma = 1 \), the filtered signal is identical to the original speech signal. On the other hand, \( \gamma = 0 \) makes the filtered signal equal to the LP-residual of \( s(n) \). It is not difficult to see that \( S_f(z) \) is the interpolated spectrum between the original and residual spectra for \( 0 < \gamma < 1 \).

### 5.5.4 Experimental Results

An objective test was conducted to evaluate the performance of the proposed PDAs in terms of pitch error rates \( (E_p) \). A test speech material of duration 114-sec was used throughout the experiments. It is sampled at 8-kHz, filtered through the modified intermediate response system [105] which simulates the analog telephone transmission system characterising a band-pass filter, and then mixed with noise signals. The speech material was composed of 58-sec male and 56-sec female speech each of which was uttered by eight speakers. The reference pitch periods of the speech signals were manually marked every 10-ms frame.

Throughout the experiments, the range of the pitch search was limited between 15 and 150 samples. Spectral analysis is conducted with the 240-point Hamming window and 256-point FFT with 16-sample zero-padding. In calculation of the TA, rectangular or tapered window with the duration 30-ms (240-sample) is applied to the input signals. The Hamming window is selected for the tapering effect.
Pitch error decisions were checked each frame by comparing a detected pitch period with the reference. A frame was classified into error if the absolute difference between the reference and the test pitch periods was more than 1-ms (8-sample) as used in [92]. Extra algorithms, such as pitch tracking using the pitch history of the past frames, were not incorporated in order to evaluate just the main algorithmic contributions.

Throughout the experiments, the voicing decision is made manually in order to evaluate the performance of the pitch estimation methods only. Aperiodic and feeble pulse regions are also included in voiced regions in consideration of its importance in speech quality. Pitch detection over those areas could increase the overall number of pitch errors.

5.5.4.1 Analysis of the STA Weighting Factor

The effect of the STA rate $\alpha$ in terms of $E_p$ is shown in Fig. 5.5. The results show that the STA gives clearly improved performance compared with the TA and the SA, corresponding to $\alpha = 1$ and $\alpha = 0$, respectively. The lowest $E_p$ was obtained when $\alpha = 0.5$ for both the female and male speeches, typically characterising high and low-pitched signals, respectively.

5.5.4.2 Analysis of the SS-SA Weighting Factor

The weighting factor $\beta$ of SS-SA in (5.42) was analysed changing $\beta$ between 0 and 1, and then the results are shown in Fig. 5.6. As observed in the analysis of the STA rate, the SS-SA also exhibits much less $E_p$ in comparison with those of the SS and the SA, corresponding to $\beta = 1$ and $\beta = 0$, respectively. Additionally, the lowest $E_p$ values were obtained when $\beta = 0.1$ for the female speech and $\beta = 0.3$ for the male speech, which mean that the optimum $\beta$ differs slightly depending on the pitch period of the signal. A higher performance can be achieved by weighting more to the SA during higher-pitched speech, but less during lower-pitched speech.
Figure 5.5: Analysis of the effect of the STA weighting factor $\alpha$ in terms of the pitch error rate. Here, the formant weighting factor $\gamma$ is 0.9.
Figure 5.6: Analysis of the effect of the SS-SA weighting factor $\beta$ in terms of the pitch error rate. Here, the formant weighting factor $\gamma$ is 0.9.
5.5. Pitch Estimation using Spectral Autocorrelation

5.5.4.3 Analysis of the Formant Weighting Factor

The effect of the formant weighting factor $\gamma$ in (5.43) was observed over the STA and SS-SA-based PDAs, and the results are shown in Fig. 5.7. From the experiment, it was found that the rate around $0.7 \sim 0.9$ exhibits satisfactory performances.

5.5.4.4 Performance Evaluation under Background Noise Environments

Experiments under noise environments were conducted using two kinds of noise sources, vehicle and babble, with various SNRs such as 5, 10, 15, 20, 25, 30, and $\infty$ dB. The STA, SS-SA and formant weighting factors, $\alpha$, $\beta$, and $\gamma$, are selected as $0.5$, $0.25$ and $0.9$, respectively. The performances of the STA and SS-SA-based PDAs were compared with those of the TA, weighted TA (WTA), SS, and weighted SS (WSS).
In WTA and WSS, the weight $W(\tau)$ for a pitch candidate period $\tau$ is defined depending on the pitch ranges in Table 5.2, as

$$W(\tau) = \begin{cases} 1, & \text{if } \tau \in \text{low-pitched} \\ \theta, & \text{if } \tau \in \text{mid-pitched} \\ \theta^2, & \text{if } \tau \in \text{high-pitched} \end{cases}$$

(5.44)

where $\theta$ is the weighting factor. The WTA and WSS metrics are finally defined by multiplying $W(\tau)$ to $R_\beta(\tau)$ in (5.36) and $\Gamma_\varphi(\tau)$ in (5.41), respectively. In WTA, the pitch error rate depending on $\theta$ is shown in Fig. 5.8. Based on the experiment, $\theta = 0.85$ is selected for the WTA with the rectangular window. In the case of TA, when $\theta = 0$, the tapered window produces much superior performance compared with the rectangular one. In the same way, the weight for the WSS is also selected as 0.85.

Experimental results over the various noise sources with levels are shown in Table 5.3 and 5.4. The results show that the STA produces performances superior to the WTA.
as well as the TA regardless of the noise sources and levels. The SS-SA-based PDA also produces performances better than the SS and WSS-based ones.

Examples of pitch contours of the various PDAs were illustrated in Fig. 5.9, 5.10, 5.11, and 5.12, in which the rectangular window is applied to the TA and STA. It shows that pitch errors in strongly voiced regions are reduced considerably in the proposed PDAs. Most of the errors of the proposed methods were caused at speech onset and offset regions characterising quite feeble and irregular pitch pulse sequences. Absence of at least strong pulses around the centre of the window lead to main performance degradation.
5.5. *Pitch Estimation using Spectral Autocorrelation*

Table 5.3: Comparison of pitch error rates among the STA, WTA, TA, SS-SA, WSS, and SS -based PDAs under vehicle noise environments.

**a) Male speech**

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Pitch error rate (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rectangular window</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STA</td>
<td>WTA</td>
<td>TA</td>
<td>STA</td>
<td>WTA</td>
<td>TA</td>
<td>SS-SA</td>
</tr>
<tr>
<td>5</td>
<td>23.76</td>
<td>31.68</td>
<td>38.90</td>
<td>28.52</td>
<td>39.23</td>
<td>32.81</td>
<td>23.62</td>
</tr>
<tr>
<td>10</td>
<td>14.66</td>
<td>21.97</td>
<td>29.42</td>
<td>19.09</td>
<td>27.72</td>
<td>23.43</td>
<td>15.61</td>
</tr>
<tr>
<td>15</td>
<td>9.57</td>
<td>16.12</td>
<td>23.29</td>
<td>13.06</td>
<td>21.40</td>
<td>17.82</td>
<td>11.41</td>
</tr>
<tr>
<td>20</td>
<td>7.69</td>
<td>13.77</td>
<td>20.74</td>
<td>10.75</td>
<td>18.81</td>
<td>15.70</td>
<td>8.77</td>
</tr>
<tr>
<td>30</td>
<td>5.24</td>
<td>11.83</td>
<td>18.62</td>
<td>8.30</td>
<td>15.98</td>
<td>13.06</td>
<td>7.21</td>
</tr>
<tr>
<td>∞</td>
<td>5.33</td>
<td>11.83</td>
<td>18.20</td>
<td>7.55</td>
<td>14.85</td>
<td>12.64</td>
<td>7.02</td>
</tr>
</tbody>
</table>

**b) Female speech**

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Pitch error rate (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rectangular window</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STA</td>
<td>WTA</td>
<td>TA</td>
<td>STA</td>
<td>WTA</td>
<td>TA</td>
<td>SS-SA</td>
</tr>
<tr>
<td>5</td>
<td>24.57</td>
<td>27.73</td>
<td>53.42</td>
<td>19.82</td>
<td>14.54</td>
<td>17.06</td>
<td>32.20</td>
</tr>
<tr>
<td>10</td>
<td>12.82</td>
<td>16.26</td>
<td>40.67</td>
<td>8.79</td>
<td>8.31</td>
<td>9.23</td>
<td>16.58</td>
</tr>
<tr>
<td>15</td>
<td>7.55</td>
<td>10.47</td>
<td>30.64</td>
<td>5.83</td>
<td>5.43</td>
<td>6.03</td>
<td>10.07</td>
</tr>
<tr>
<td>20</td>
<td>5.23</td>
<td>7.71</td>
<td>25.17</td>
<td>4.12</td>
<td>4.04</td>
<td>4.75</td>
<td>7.91</td>
</tr>
<tr>
<td>25</td>
<td>4.16</td>
<td>7.03</td>
<td>22.53</td>
<td>3.68</td>
<td>3.92</td>
<td>4.19</td>
<td>6.79</td>
</tr>
<tr>
<td>∞</td>
<td>3.52</td>
<td>6.41</td>
<td>19.91</td>
<td>2.75</td>
<td>3.27</td>
<td>3.73</td>
<td>5.91</td>
</tr>
</tbody>
</table>
5.5. Pitch Estimation using Spectral Autocorrelation

Table 5.4: Comparison of pitch error rates among the STA, WTA, TA, SS-SA, WSS, and SS-based PDAs under babble noise environments.

(a) Male speech

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Pitch error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rectangular window</td>
</tr>
<tr>
<td>5</td>
<td>30.41 36.68 43.80</td>
</tr>
<tr>
<td>10</td>
<td>18.25 24.71 32.48</td>
</tr>
<tr>
<td>15</td>
<td>11.60 17.54 24.33</td>
</tr>
<tr>
<td>20</td>
<td>8.68 14.47 20.70</td>
</tr>
<tr>
<td>25</td>
<td>6.69 12.68 18.95</td>
</tr>
<tr>
<td>30</td>
<td>5.99 11.98 18.82</td>
</tr>
<tr>
<td>∞</td>
<td>5.33 11.83 18.20</td>
</tr>
</tbody>
</table>

(b) Female speech

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Pitch error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rectangular window</td>
</tr>
<tr>
<td>5</td>
<td>21.42 27.81 51.80</td>
</tr>
<tr>
<td>10</td>
<td>12.55 15.78 39.13</td>
</tr>
<tr>
<td>15</td>
<td>7.80 10.19 30.31</td>
</tr>
<tr>
<td>20</td>
<td>5.44 8.07 24.70</td>
</tr>
<tr>
<td>25</td>
<td>4.71 7.11 21.98</td>
</tr>
<tr>
<td>30</td>
<td>4.05 6.91 20.74</td>
</tr>
<tr>
<td>∞</td>
<td>3.52 6.41 19.91</td>
</tr>
</tbody>
</table>
Figure 5.9: Comparison of pitch contours of PDAs for (a) noise-free female speech, (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs.
Figure 5.10: Comparison of pitch contours of PDAs for (a) noise-free male speech (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs.
5.5. Pitch Estimation using Spectral Autocorrelation

Figure 5.11: Comparison of pitch contours of PDAs for (a) female speech corrupted by vehicle noise of 15 SNR dB. (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA -based PDAs.
Figure 5.12: Comparison of pitch contours of PDAs for (a) male speech corrupted by vehicle noise of 15 SNR dB. (b) Reference; (c) TA, (d) WTA, (e) STA, (f) SS, (g) WSS, and (h) SS-SA-based PDAs.
Table 5.5: Comparison of the computational complexity.

<table>
<thead>
<tr>
<th>Computational complexity (MIPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No decimation</td>
</tr>
<tr>
<td>STA</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>Decimation</td>
</tr>
<tr>
<td>Format weighting</td>
</tr>
<tr>
<td>TA</td>
</tr>
<tr>
<td>FFT</td>
</tr>
<tr>
<td>SA</td>
</tr>
<tr>
<td>Lag search</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

5.5.4.5 Computational Complexity

In applications, such as speech coding, the computational complexity of the pitch estimation method is a great concern. Thus, the complexity of the proposed pitch estimation method is evaluated in terms of the number of fixed-point instructions. For each algorithmic sub-block comprising the STA and WTA methods, the complexities are measured and then shown in Table 5.5. The main components which require high complexity are the TA and formant weighting which are common algorithmic sub-blocks to both STA and WTA. The SA itself does not require huge computation compared to TA because a large portion of the computation is redundant due to the limit in the resolution of \( \omega_p \), as shown in Fig. 5.13. For the case of \( M = 256 \), the number of required lags for SA calculation over the pitch search range is 16. Moreover, the computational complexity can be reduced significantly by using decimated signals because voiced signals exhibit strong harmonics at lower frequencies. The required computation for half decimated signals is also shown in Table 5.5.
Figure 5.13: The relationship between the spectral lag $\omega_r$ (the number of frequency bins) and the temporal lag $\tau$ (the number of time-domain samples) for the DFT size $M = 256$. 
5.6 Predictive and Mel-scale Binary Vector Quantisation

5.6.1 Introduction

Quantisation of the spectral magnitudes of the linear predictive (LP) residual is one of the crucial parts in sinusoidal speech coders as the accuracy substantially affects the speech quality [78][89][106]. The problem inherent in this task is that the dimension of the spectral magnitudes varies from frame to frame depending on the pitch, typically ranging between 10 and 80.

From psychoacoustic evidences, it is well known that the frequency resolutions of higher frequencies is worse than those of lower frequencies [107]. Thus mel-scale, which warps the frequency range in the logarithmic scale, can be introduced to quantise the spectral magnitudes efficiently. Moreover, the LP-envelope exhibits a significant correlation between successive frames [86], especially on steady voiced regions, which leads to the high correlation of the spectral magnitudes of the LP-residual. In other words, the inter-frame correlation can be exploited for the efficient quantisation of the spectral magnitudes.

Hence, this section proposes a novel method for the quantisation of the spectral magnitudes by predictive and mel-scale binary vector quantisation (PMBVQ), which is effective to attain high spectral accuracy as well as to quantise the magnitude with very low computational complexity. The overall structure of the quantiser is shown in Figure 5.14.

5.6.2 Distortion Measure

The spectral magnitude $y$ can be modelled by the product of LP-spectral envelope $H$ and excitation spectrum $x$ as

$$ y = Hx $$

(5.45)

where the diagonal elements of $H$ are the magnitude of the frequency response of the LP-coefficient of the synthesis filter; and the dimensions of $H$, $y$ and $x$ are $K \times K$,
5.6. Predictive and Mel-scale Binary Vector Quantisation

Figure 5.14: Block diagram of predictive and mel-scale binary vector quantiser.

$K \times 1$ and $K \times 1$, respectively, in which $K$ is determined by the fundamental frequency $\omega_0$, i.e., $K = \lfloor \pi/\omega_0 \rfloor$.

A weighted distortion measure is defined, so as to quantise model parameters efficiently, as

$$\varepsilon = \left\| WH(x - g_p x_p - g_c x_c) \right\|^2$$

(5.46)

where the excitation vector $x$ is decomposed into the predictive and linear-scale code vectors, $x_p$ and $x_c$, whose gains are $g_p$ and $g_c$, respectively. The diagonal elements of the $K \times K$ matrix $W$ are the magnitude of the frequency response of the Atal's perceptual weighting filter as

$$A_{ip}(z) = e^{jk\omega_0}, 1 \leq k \leq K,$$

(5.47)

where $A_{ip}(z)$ is $z$-transform representation of the LP inverse filter. The weighting factors are selected through experiments, as $\gamma_1 = 1.0$ and $\gamma_2 = 0.9$, which emphasize the spectral valleys rather than the peaks.

From minimization procedure of the error measure (5.46) in closed-loop, it is possible to quantise the parameters of the excitation spectrum, such as $g_p$, $g_c$, and $x_c$. The model parameters of the decomposed excitations are optimised sequentially to reduce computational complexity. In other words, the gain of the predictive excitation is
5.6. Predictive and Mel-scale Binary Vector Quantisation

derived firstly based on the assumption that the gain of the code excitation is zero, and then an optimum code excitation with its gain is calculated.

5.6.3 Predictive Quantisation

From observation of LP-residual spectra, it is found that the spectra change slowly from frame to frame. In other words, there is high statistical correlation between consecutive residual spectra at the point of entropy coding. Hence, the predictive coding in conjunction with its residual spectrum coding could be useful for reducing the number of bits to represent the spectral magnitude, rather than quantising the residual magnitude directly or increasing the LP-model order.

The predictive excitation spectrum $x_p$ is derived from the previous excitation spectrum $\hat{x}^{(t-1)}$. The dimension of spectral vectors is, however, variable from frame to frame, thus the predictive vector $x_p$ of $K$-dimension is obtained by linearly warping the synthesized excitation vector of the previous frame $\hat{x}^{(t-1)}$ of $K^{(t-1)}$-dimension as

$$x_p(k) = \hat{x}^{(t-1)} \left( \left\lfloor \frac{K^{(t-1)}}{K} k + 0.5 \right\rfloor \right), 1 \leq k \leq K,$$

where $x_p(k)$ is the $k$th spectral component of $x_p$. The predictive gain $g_p$ can be obtained by setting the first order derivative of the error measure (5.46) to zero with respect to the predictive gain, as

$$g_p = \frac{x_p^T H^T W^T \hat{W} x_p}{x_p^T H^T W^T \hat{W} x_p}.$$  \hspace{1cm} (5.49)

To reconstruct the predictive spectrum $x_p$ at the decoder, bit-transmission is not required as it is obtained from the reconstructed excitation spectrum of the previous frame, $\hat{x}^{(t-1)}$.

5.6.4 Mel-Scale Binary Vector Quantisation

The residual spectrum, defined the difference between the original and the LP-and-predictive envelopes, is considered for spectral compensation. In quantisation of the residual spectrum, there are two problems: 1) the dimension of the residual spectrum
is variable, and 2) the spectral dimension can be large, thus a number of bits and codebook space are required to directly quantise the spectrum.

Thus, this thesis proposes a mel-scale binary codebook (MBCB), which represents the residual spectrum of a variable and large dimensions with a code vector of a fixed and low one. Non-linear frequency scaling, such as mel, to the frequency axis is performed since the harmonic bins of the lower frequencies are perceptually more important than those of the higher ones. This scheme splits harmonic frequency bins into a fixed number of sub-bands, in the logarithmic scale. Finally, each sub-band is quantised by a binary value. Hence, the $k$th component of the linear-scale code vector $x_c(k)$ can be obtained from the $m$th element of the mel-scale code vector $c(m)$ in MBCB by mel-to-linear transformation as

$$x_c(k) = c \left( \left( M - 1 \right) \log_2 \left( \frac{k}{K} \right) + 1 \right), \text{ for } 1 \leq k \leq K,$$

where $M$ is the dimension of the mel-scale code vector $c$. The conversion changes the number of frequency bins for each sub-band depending on the pitch, and generates a variable-dimension vector from the fixed-dimension code vector. The fixed dimension of the code vector is relatively low, e.g. 10, 12, or 14, compared with the number of harmonics, typically ranging between 10 and 80. Hence, it is possible to generate $x_c$ of $K$-dimension from the code vector $c$ of $M$-dimension by the mel-to-linear transform.

The optimum code vector $c^*$ can be obtained through analysis-by-synthesis procedure in the frequency domain as

$$c^* = \text{arg} \max_{c \in \Omega} \frac{(x^T - g_c x_p^T)H^T W^T W H x_c)^2}{x_c^T H^T W^T W H x_c},$$

where $\Omega$ means the set of code vectors in MBCB, composed of $2^M$ code vectors, and the linear-scale code vector $x_c$ can be obtained from the mel-scale code vector $c$ by (5.50). The optimum gain $g_c$ for the optimum code vector $x_c^*$ can be obtained as

$$g_c = \frac{(x_c^T - g_c x_p^T)H^T W^T W H x_c}{x_c^T H^T W^T W H x_c^*}.$$

### 5.6.5 Complexity Optimisation

Most of the computation of PMBVQ is caused by the closed-loop computation in (5.51) to find the optimum code vector $c^*$ together with its gain $g_c$. If the binary code value
5.6. Predictive and Mel-scale Binary Vector Quantisation

$x_c(k)$ has a same absolute value, $+1$ or $-1$, the closed-loop search can be rewritten as

$$c^* = \arg\max_{c \in \Omega} ((x^T - g_p x_p^T) H^T W^T W x_c)^2,$$

(5.53)
since $x_c^2(k) = 1$, for $1 \leq k \leq K$. Subsequently $x_c^2 H^T W^T W x_c$ in (5.51) becomes a constant. Furthermore, (5.53) can be simplified as

$$c^* = \arg\max_{c \in \Omega} (d^T x_c)^2$$

(5.54)
where $d = H^T W^T W (x - g_p x_p)$. The maximum value of (5.54) is limited as

$$\left( \sum_{m=0}^{M-1} \left\{ c(m) \sum_{k=l_m}^{u_m} d(k) \right\} \right)^2 \leq \left( \sum_{m=0}^{M-1} \sum_{k=l_m}^{u_m} d(k) \right)^2$$

(5.55)
where $c(m) = \pm 1$, and $d(k)$ is the $k$th component of the vector $d$. In (5.55), $l_m$ and $u_m$ are the lower and the upper harmonic bounds of the $m$th sub-band, $[(K-1)(2^m - 1) + 1.5]$ and $[(K-1)(2^m - 1) + 0.5]$, respectively. Hence, the optimum code vector $c^*$ can be obtained efficiently by way of open-loop search as

$$c^*(m) = \begin{cases} 1 & \text{if } \sum_{k=l_m}^{u_m} d(k) > 0, \\ -1 & \text{elsewhere,} \end{cases}$$

(5.56)
or

$$c^*(m) = \begin{cases} -1 & \text{if } \sum_{k=l_m}^{u_m} d(k) > 0, \\ 1 & \text{elsewhere,} \end{cases}$$

(5.57)
where $c^*(m)$ is the $m$th component of the optimum code vector $c^*$, for $0 \leq m \leq M - 1$.

PMBVQ with complexity optimisation performs magnitude quantisation with a very low complexity in terms of the required memory and time, since it quantises the spectral magnitude without a trained codebook and finds the optimum code vector with open-loop search.

5.6.6 Experimental Results

The performance of PMBVQ is evaluated in terms of the weighted signal to noise ratio (WSNR) in the spectral domain, defined

$$\text{WSNR(dB)} = 10 \log \frac{y^T W^T W y}{(y - \hat{y})^T W^T W (y - \hat{y})}$$

(5.58)
Table 5.6: Comparison of WSNR(dB) between the PMBVQ and the LP-only methods.

<table>
<thead>
<tr>
<th>LP order</th>
<th>LP-only 8-dimension</th>
<th>LP-only 10-dimension</th>
<th>PMBVQ 12-dimension</th>
<th>PMBVQ 14-dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>8.9</td>
<td>10.5</td>
<td>11.0</td>
<td>11.5</td>
</tr>
<tr>
<td>8</td>
<td>10.0</td>
<td>11.5</td>
<td>11.9</td>
<td>12.4</td>
</tr>
<tr>
<td>10</td>
<td>10.5</td>
<td>11.9</td>
<td>12.4</td>
<td>12.9</td>
</tr>
<tr>
<td>14</td>
<td>11.3</td>
<td>12.7</td>
<td>13.1</td>
<td>13.6</td>
</tr>
<tr>
<td>18</td>
<td>12.0</td>
<td>13.1</td>
<td>13.5</td>
<td>14.0</td>
</tr>
<tr>
<td>22</td>
<td>12.5</td>
<td>13.6</td>
<td>14.0</td>
<td>13.4</td>
</tr>
<tr>
<td>26</td>
<td>13.0</td>
<td>13.9</td>
<td>14.3</td>
<td>14.6</td>
</tr>
</tbody>
</table>

where $\hat{y} = g_p H x_p + g_c H x_c$.

The speech material of 96 sec sampled at 8 kHz is collected for the experiments. It is analysed every 10 ms, and finally 9600 spectral vectors are obtained. From the experiment, it is apparent that the increase of the dimension of the code vector is more effective than the increase of the order of the LP-model as shown in Table 5.6, because MBVQ requires one bit per dimension, however LP quantisation typically needs about 2 ~ 3 bits per LP-order. Furthermore, MBVQ does not require a codebook space to keep the code vectors while LP quantisation may require a large one. The relationship, obtaining similar WSNR, between the dimension of PMBVQ and the order of the LP-model is shown in Figure 5.15. To be specific, 12-dimension PMBVQ with the 10th order LP-model is comparable to the 26th order LP-only method.

Additionally, the performances of PMBVQ and the higher order LP-model depending on pitch periods are compared as shown in Figure 5.16. In the case of PMBVQ, the order of the LP-model and the dimension of the fixed code vector are fixed to 10 and 12, respectively. On the other hand, the model order of the LP-only method is 26. This experiment obviously shows that PMBVQ in conjunction with the low-order LP-model is superior to the high order LP-model for voiced speech, however a little degradation in WSNR for unvoiced (UV) speech. It is a widely-known fact that spectral degradation
is not perceptually significant for unvoiced speech, but important for voiced speech.

Figure 5.16 also shows that an increase of the pitch period leads to a decrease of WSNR, which is a natural result in that the lower-pitched voice has a larger number of harmonic bins for each spectral sub-band. In sinusoidal speech coding, however, it is reported that the spectral magnitude is perceptually more effective for high-pitched sound rather than low-pitched one [108]. From the informal listening test, it is also confirmed that low-pitched voice is less sensitive to spectral distortion compared to high-pitched one.

5.7 Switched Predictive Mel-scale-based Vector Quantisation

5.7.1 Introduction

This section presents another efficient spectral-magnitude quantisation method, switched-predictive mel-scale-based vector quantisation (SP-MVQ), based on the perceptual and statistical redundancies of the LP-residual magnitude-spectra.

Subsection 5.7.2 describes the detailed structure of the SP-MVQ method. Subsection
Figure 5.16: Performance comparison between PMBVQ and a high order LP-model against the pitch. The order of the high order LP-model is 26. The vector dimension and LP-order for PMBVQ are 12 and 10, respectively.

Figure 5.17: Distribution of the pitch of the speech material used in the test.
5.7.3 presents the results from performance evaluation using an objective measure. The effect of the subjective speech-quality of SP-MVQ is also given by incorporating the magnitude quantiser in a sinusoidal speech coder.

### 5.7.2 Switched-Predictive Mel-scale-based Vector Quantisation

The spectral magnitude vector \( y \) estimated at the pitch harmonic frequencies of speech signals can be decomposed into the root mean square (RMS) gain \( g \), the LP-spectral envelope \( H \), and the normalised spectral-magnitude vector \( x \) of the LP-residual signal, as

\[
y = gHx
\]

where the dimensions of \( y, H, \) and \( x \) are \( K \times 1, K \times K, \) and \( K \times 1, \) respectively. The dimension \( K \) can vary from frame to frame, and can be determined by the pitch \( \omega_0 \) of the frame being analysed, such that \( K = \lfloor \pi / \omega_0 \rfloor \). The LP-spectral envelope \( H \) is a diagonal matrix whose elements are the spectral amplitudes of the LP synthesis filter.

Taking into account the perceptual preferences of the human auditory system, the variable dimension spectral vector \( x \) is converted into a fixed-dimension vector \( z, M \times 1, \) using mel-scale-based warping as [106]

\[
z(m) = \sqrt{\frac{1}{u_m - l_m + 1} \sum_{k=l_m}^{u_m} x^2(k)}
\]

where \( x(k) \) and \( z(m) \) denote the \( k \)th and the \( m \)th elements of the vector \( x \) and \( z \), respectively; and \( l_m \) and \( u_m \) denote the lower and upper harmonic bounds of the \( m \)th spectral band, \( [K \frac{m-0.5}{M} - 0.5] \) and \( [K \frac{m+1}{M} - 1.5] \) (but \( u_{M-1} = K - 1 \)), respectively.

From investigation of the LP-residual magnitude spectra, it is observed that there is a considerable redundancy between the consecutive spectra, as shown in Figure 5.18. Thus, the fixed-dimension LP-residual vector is decomposed into the predicted vector \( z_p \) and the prediction residual vector \( z_c \). In other words, the quantised vector \( \hat{z} \) is obtained as follows:

\[
\hat{z} = z_p + \hat{z}_c
\]
Figure 5.18: Analysis of the inter-frame correlation of LP-residual magnitude-spectra: (a) waveform of a female speech segment, (b) developed, every 10-ms, LP-residual magnitude-spectra of the boxed speech in (a), (c) normalised correlation ratio of the frame-to-frame harmonic sequence in (b), and (d) overlapped display of (b).
5.7. Switched Predictive Mel-scale-based Vector Quantisation

where \( \hat{z}_c \) is the quantised vector of \( z_c \). The predicted vector is obtained using a first-order auto-regressive method as

\[
z_p = \Phi (\hat{z}_d - z_m) + z_m
\]  

(5.62)

where \( \hat{z}_d \) means the most recently quantised vector of \( \hat{z} \); and \( \Phi \) and \( z_m \) denote the \( M \times M \) prediction matrix and the \( M \times 1 \) mean vector, respectively. It is also possible to generalise the prediction scheme in (5.62) using higher-order predictors. In our experiment, however, each component of \( z_m \) is fixed to 1.0, and \( \Phi \) is the diagonal matrix in which each component is obtained through the correlation analysis of 10 ms block shift. The prediction residual vector \( z_c \) is quantised by a typical vector quantiser, such as a multi-stage VQ [83]. The quantisation becomes memoryless mel-scale-based vector quantisation (ML-MVQ) if all prediction coefficients are equal to zero, and auto-regressive predictive MVQ (P-MVQ) otherwise.

The predictive quantisation is effective in very steady regions, but may cause an increase in the spectral distortion especially in transitional regions. To circumvent this problem a switching scheme between the predictive and the memoryless quantisers can be considered. Thus SP-MVQ is designed, consisting of P-MVQ and ML-MVQ, in order to give optimal performance over both steady and transitional regions through a switching scheme. The optimal switch-mode \( \sigma^* \) and code vector index \( \nu^* \) of SP-MVQ are determined by an analysis-by-synthesis procedure in terms of the weighted

Figure 5.19: Block diagram of SP-MVQ encoder.
5.7. Switched Predictive Mel-scale-based Vector Quantisation

Distortion measure, as

\[
(\sigma^*, \nu^*) = \arg \min_{\sigma, \nu} \| W(y - gH\hat{x}_{\sigma, \nu}) \|^2
\]  

(5.63)

where \( \sigma \) means the switch mode out of ML-MVQ and P-MVQ, and \( \nu \) means the index of the code vector. The reconstructed linear-scale magnitude vector \( \hat{x} \) can be attained from \( \hat{x} \) by the inverse warping procedure as

\[
\hat{x}(k) = \hat{x} \left( \left( M - 1 \right) \log_2 \left( \frac{k - 1}{K - 1} + 1 \right) \right)
\]  

(5.64)

where \( \hat{x}(k) \) and \( \hat{x}(m) \) denote the \( k \)th and \( m \)th elements of \( \hat{x} \) and \( \hat{z} \), respectively. Here, the reconstructed values of the linear spectral components in a mel-band are identical. The \( K \times K \) diagonal matrix \( W \) is applied, for shaping quantisation noise by emphasising the spectral valleys rather than peaks, as

\[
w_{kk} = \left| \frac{A_{lp}(x/\gamma_1)}{A_{lp}(x/\gamma_2)} \right|, x = e^{j\omega_0}, 1 \leq k \leq K
\]  

(5.65)

where \( w_{kk} \) denotes the \( k \)th diagonal element of \( W \), and \( A_{lp}(x) \) is the LP inverse filter. The weighting factors, \( \gamma_1 \) and \( \gamma_2 \), are determined through intensive experiments as 1.0 and 0.9, respectively.

The structure of the SP-MVQ encoder is illustrated in Figure 5.19, in which the minimisation and the mel-to-linear transform procedures are carried using (5.63) and (5.64), respectively. The decoder of SP-MVQ is the subset of the encoder, due to the analysis-by-synthesis structure of the encoding process.

5.7.3 Experimental Results

The performance of the proposed vector quantiser is evaluated by an objective measure in terms of the weighted SNR (WSNR), defined

\[
\text{WSNR} = \frac{x^T H^T W^T W x}{(x - \hat{x})^T H^T W^T W (x - \hat{x})}
\]  

(5.66)

in which \( W \) is the diagonal weight matrix as defined in (5.65). In the experiment, 115,200 spectral vectors with 10 ms block shift are collected from NTT-AT speech.

\( ^{1} \)The VQ switch mode \( \sigma \) and the vector index \( \nu \) are omitted for the simplicity of the description
5.7. Switched Predictive Mel-scale-based Vector Quantisation

Figure 5.20: Performance comparison between ML-MVQ, P-MVQ, and SP-MVQ, of which the numbers of allocated bits are 14, 14, and 13+1(VQ switch-mode), respectively.
Table 5.7: Bits allocation of the SB-LPC vocoder with SP-MVQ. The frame size is 20 ms, and the frame is split into two subframes.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Bits/frame (subframe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP-coefficient</td>
<td>23</td>
</tr>
<tr>
<td>Pitch</td>
<td>12 (5 + 7)</td>
</tr>
<tr>
<td>Voicing</td>
<td>6 (3 + 3)</td>
</tr>
<tr>
<td>RMS-Gain</td>
<td>10 (5 + 5)</td>
</tr>
<tr>
<td>Spectral magnitude</td>
<td>28 (14 + 14)</td>
</tr>
<tr>
<td>Parity</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

The effect to the subjective speech-quality of SP-MVQ is investigated by incorporating the quantiser in the split-band LPC (SB-LPC) vocoder [109]. The model parameters of the SB-LPC vocoder consist of LP-coefficient, pitch, voicing level, gain, spectral magnitudes, and parity bit. The allocated bits for these parameters, to achieve a 4 kbit/s vocoder over 20 ms frame, are shown in Table 5.7. The fixed dimension of the spectral magnitudes is 24. For the listening test, eight sentences uttered by four-male and four-female speakers are collected, filtered by m-IRS, and processed by both 4 kbit/s SB-LPC and 8 kbit/s G.729 [12] speech coders. Eight listeners are asked to indicate their preferences out of the randomised pairs of the processed speech samples.

The subjective test results are shown in Table 5.8. Table 5.8 (a) indicates that the SB-LPC coder with SP-MVQ gives speech quality slightly better than G.729 for the female database. The LP coefficients are not quantised in order to reduce the effects of the other parameters. The result is illustrated in Figure 5.20, which shows that SP-MVQ gives performances superior to ML-MVQ and slightly superior to P-MVQ. Decorrelating the predictive term from the LP-residual vector leads to the improved quantisation performance. Note that the switching scheme is useful for preventing the propagation of errors due to channel noises by limiting the number of consecutive selections of P-MVQ. Hence, SP-MVQ is more desirable than P-MVQ, especially in noisy channel environments.
speech, but slightly worse than for the male speech. This is due to the fact that male voices are normally lower-pitched in comparison with female voices. In other words, the number of harmonic components per mel-band of the lower-pitched voice is larger than that of the higher-pitched voice. Thus, the spectral distortion due to mel-scale conversion causes the speech quality degradation. Preference test results for the original and the flat magnitude schemes are also given in Table 5.8 (b) and (c), to confirm the importance of the spectral magnitude accuracy in sinusoidal speech coding. This shows that the original magnitude scheme produces highly desirable speech quality while the flat magnitude scheme produces much worse quality than G.729. The mel-scale dimension and the number of bits allocated are selected in consideration of developing the low bit-rate SB-LPC coder.

5.8 Summary

In the chapter, we have reviewed low bit-rate speech coding techniques, and proposed three new algorithms for improving the performance of speech coders. It has been reviewed that the sinusoidal speech model has good features for lowering the speech coding bit-rate by removing phase redundancy in speech signal. Fundamental techniques for designing high quality sinusoidal speech coders also have been introduced.

The STA-based PDA has been designed for robust pitch estimation by exploiting the dual property between TA and SA. Additionally, SA has also been applied to SS since the duality is valid between SS and SA, and then the SS-SA-based PDA has been designed. Performance evaluation in terms of the pitch error rate has confirmed that both STA and SS-SA PDAs produce performances much superior to conventional PDAs using WTA and WSS.

PMBVQ has been proposed for efficient quantisation of the magnitude spectrum of the LP-residual signal. The predictive coding scheme in PMBVQ removes the redundancy in LP-residual spectra between consecutive frames. The predictive coding error is compensated by MBVQ, which converts spectral harmonics of variable dimensions into spectral sub-bands of a fixed dimension using mel-scale warping. The effectiveness of the proposed coding method is evaluated in terms of WSNR. From the experiment,
Table 5.8: Preference test results.

(a) SB-LPC with SP-MVQ

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-LPC</td>
<td>26.6</td>
<td>25.0</td>
<td>28.1</td>
</tr>
<tr>
<td>G.729</td>
<td>28.1</td>
<td>34.4</td>
<td>21.9</td>
</tr>
<tr>
<td>No preference</td>
<td>45.3</td>
<td>40.6</td>
<td>50.0</td>
</tr>
</tbody>
</table>

(b) SB-LPC with original magnitude

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-LPC</td>
<td>33.8</td>
<td>34.4</td>
<td>37.4</td>
</tr>
<tr>
<td>G.729</td>
<td>25.0</td>
<td>12.5</td>
<td>31.3</td>
</tr>
<tr>
<td>No preference</td>
<td>41.2</td>
<td>53.1</td>
<td>31.3</td>
</tr>
</tbody>
</table>

(c) SB-LPC with flat magnitude

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB-LPC</td>
<td>15.6</td>
<td>18.8</td>
<td>12.5</td>
</tr>
<tr>
<td>G.729</td>
<td>62.5</td>
<td>53.1</td>
<td>71.7</td>
</tr>
<tr>
<td>No preference</td>
<td>21.9</td>
<td>28.1</td>
<td>16.1</td>
</tr>
</tbody>
</table>
it has been confirmed that the proposed method can represent spectral magnitudes with a high spectral accuracy in comparison with high-order LPC. Additionally, it can perform spectral magnitude quantisation with a very low computational complexity and memory requirement.

Another efficient spectral-magnitude quantisation method, SP-MVQ, has been proposed. It features mel-scale-based dimension conversion to handle the variable dimension spectral vector and the switched-predictive scheme to exploit the inter-frame correlation of the spectral vectors. The result of the subjective speech-quality test has confirmed that SP-MVQ is highly desirable as a spectral magnitude quantisation method for sinusoidal speech coders. The proposed quantisation schemes can also be applied to waveform interpolation (WI) coders, which have been widely investigating for low bit-rate speech coders. In the WI coder, quantisation of pitch cycle waveforms in which the dimension changes from frame to frame is a key component in order to produce high-quality speech. Thus, it is highly expected that the proposed method will be useful for the efficient quantisation of the pitch cycle waveform of the WI-based speech coder.
Chapter 6

An Enhanced Split-band LPC Vocoder

The objective of this chapter is to propose an enhanced low bit-rate speech coder with the capability of silence compression, noise reduction, and improved model parameters estimation and quantisation. Section 6.1 introduces the evolution of a low bit-rate speech coder. Section 6.2 describes the overall structure of an enhanced coder. Section 6.3 describes signal preprocessing for removal of very low frequency components. Section 6.4 describes VAD and noise reduction techniques. Section 6.5 and 6.6 present the functional descriptions of vocoder model parameters estimation and quantisation, respectively. Packetisation for transmission through telecommunication networks is shown in Section 6.7. Section 6.8 describes the speech synthesis method. The comfort noise generation and insertion schemes for handling silence frames are presented in Section 6.9. The perceptual speech-quality of the proposed coder is evaluated in Section 6.10. Finally the summary of the chapter is presented in Section 6.11.

6.1 Introduction

The sinusoidal model represents speech signals with the sum of sinusoidal terms sampled at harmonic frequencies. The representation can be enhanced by decomposing the speech into the vocal source and filter terms, and then by applying the sinusoidal model
6.2 Overall Structure of the eSB-LPC Speech Encoder and Decoder

to the vocal source signals. The reason is that the vocal source signal rather than the
speech is more suitable for representing with the zero-and-random phase model. Based
on this motivation, Yeldener et al. proposed the multi-band excited LPC (MBE-LPC)
 vocoder [72][69], which reconstructs speech signals by exciting the LP-synthesis filter
with the sinusoidal excitation source. The phase model can be simplified by treating
the phase of the residual signal rather than the complicated one of the speech.

Sinusoidal model-based coders would produce smeared sounds due to incorrect estima-
tion of the model parameters. Especially the pitch in the model parameters consider-
ably affects the speech quality. It is not difficult to see severe degradation of the speech
quality in sinusoidal speech coders when the pitch for a voiced region is not accurate.
Atkinson et al. thus proposed enhanced estimation schemes for the pitch together with
the voicing level, called the split-band LPC (SB-LPC) vocoder [81][71].

This chapter proposes a further enhanced version, named enhanced SB-LPC (eSB-
LPC) vocoder, featuring a few additional functions together with even further improved
model parameter estimation. Firstly, eSB-LPC is capable of silence compression by
means of voice activity detection, comfort noise insertion, and comfort noise generation.
Secondly, reduction of background noise from the input signal is introduced because the
synthesised speech can be easily corrupted by the noise. Thirdly, a quantisation method
for the spectral magnitudes is employed to achieve an enhanced speech-quality, because
the spectral magnitude of the LP-residual signal plays an important role in sinusoidal
speech coders as discussed in Section 5.7. Lastly, a new pitch estimation method is
employed for the robustness against various kinds and levels of noise sources.

6.2 Overall Structure of the eSB-LPC Speech Encoder
and Decoder

The overall structure of the eSB-LPC codec is shown in Figure 6.1. Moreover, further
detailed structure of the encoder is illustrated in Figure 6.2. The frame buffer is filled
with the input speech of 20 ms corresponding to 160 samples with a sampling frequency
of 8 kHz.
6.2. Overall Structure of the eSB-LPC Speech Encoder and Decoder

Figure 6.1: Block diagram of the eSB-LPC codec.

For input signals, the encoder filters out very low frequency components using a high pass filter (HPF). VAD decides the presence of speech for the input frame. Noise reduction suppresses the noise components in noisy speech, and then produces enhanced speech. Based on the sinusoidal model, the speech encoder conducts the estimation and quantisation of the speech model parameters consisting of pitch, voicing level, gain, LP coefficients, and spectral magnitudes. Finally, at the encoder, the quantised results are packed into a bit stream to transmit to the decoder. The bit stream has three packet structures depending on a decision of voice activity and comfort noise insertion (CNI). Voice-active frames are compressed at the rate of 4 kbit/s. On the other hand, voice-inactive frames are packed in either 0 or 1.7 kbit/s depending on the flag of CNI. Each algorithmic processing except LP-coefficient quantisation is carried out every subframe of 10 ms, but the LP-coefficient is quantised every frame of 20 ms. The speech decoder unpacks the received bit stream, and then reconstructs the model parameters to synthesise output speech signals. A proper synthesis mechanism is selected depending on the modes of voice activity and CNI. The input and output parameters for forward and inverse packetisation are shown in Figure 6.3.
6.2. Overall Structure of the eSB-LPC Speech Encoder and Decoder

![Block diagram of the eSB-LPC encoder.](image)

Figure 6.2: Block diagram of the eSB-LPC encoder.
6.2. Overall Structure of the eSB-LPC Speech Encoder and Decoder

![Block diagram of the forward and inverse packetisation of the eSB-LPC codec.]

Figure 6.3: Block diagram of the forward and inverse packetisation of the eSB-LPC codec.
6.3 Preprocessing

The buffered input signal is filtered through a HPF to remove any DC component and electrical power noise mixed with the speech signal. Specifically, the input signal is filtered by the second order infinite impulse response filter $H_{hp}(z)$ of the form

$$H_{hp}(z) = \frac{0.92727436 - 1.85449410z^{-1} + 0.92727436z^{-2}}{1 - 1.9059465z^{-1} + 0.9114024z^{-2}}.$$

6.4 Voice Activity Detection and Noise Reduction

VAD is conducted to decide the presence of speech in the current input frame using the smoothed likelihood ratio test-based method, proposed in Section 3.5. VAD is carried out every 10 ms with the spectral amplitudes composed of 128 bins, in which the spectral analysis is performed by FFT for the windowed speech signal of 256 samples. The final VAD for a frame of 20 ms is decided by combining the two 10-ms results. A frame is decided as voice-active if at least one subframe is detected as voice-active.

Noise reduction is conducted to enhance the speech signals against background noise by MMSE-LSA-SD, described in Sections 4.2.5 and 4.2.6, together with the mixed decision-based noise adaptation technique, proposed in Section 4.3. Noise reduction is performed every 10 ms using the FFT spectrum of 256-point.

6.5 Estimation of Model Parameters

The model parameters of the eSB-LPC vocoder are composed of pitch, voicing level, LP coefficients, gain, and spectral magnitudes.

The pitch estimation is conducted in two steps. Firstly, the integer pitch is estimated for the range between 15 and 150 based on the spectro-temporal autocorrelation method, described in Section 5.5. Secondly, the pitch is refined to a fractional level of 0.125 sample searching around the estimated integer pitch, by means of the Atkinson's pitch estimation method [71] based on the spectral analysis-by-synthesis method.
6.5. *Estimation of Model Parameters*

The voicing level is defined by a voicing cutoff frequency of which the lower and the higher bands indicate voiced and unvoiced harmonics, respectively. The cutoff frequency is estimated using the Atkinson's method \([71]\) based on the similarity between each spectral harmonic shape and the applied window. In other words, a spectral harmonic is classified as voiced if the similarity is relatively high, otherwise as unvoiced. The optimal harmonic index separating the voiced and unvoiced bands is determined by a distortion measure in \([71][81][109]\).

LP coefficients are computed based on the autocorrelation method. For an input signal \(s(n)\), the windowed signal \(s_w(n)\) is produced as

\[
s_w(n) = s(n)w_H(n), \quad \text{for } 0 \leq n \leq N - 1, \tag{6.2}
\]

where \(N = 240\), and \(w_H(n)\) is the Hamming window given

\[
w_H(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N} \right). \tag{6.3}
\]

With the autocorrelation ratio \(R(i)\), given

\[
R(i) = \sum_{n=0}^{N-i-1} s_w(n)s_w(n+i), \quad \text{for } 0 \leq i \leq p, \tag{6.4}
\]

the \(p\)th order LP coefficients \(\tilde{a}_i\) selected \(p = 10\), are derived recursively by the Levinson-Durbin algorithm described in Section 5.4. The LP synthesis filter using \(\tilde{a}_i\) characterises the all-pole model which may exhibit sharp resonances for the pole frequencies. Thus, the bandwidth of the spectral poles are expanded by modifying the LP coefficients, as

\[
a_i = \gamma^i \tilde{a}_i, \quad \text{for } 1 \leq i \leq p, \tag{6.5}
\]

where \(\gamma\) is the bandwidth expansion factor, \(0 \leq \gamma \leq 1\), selected 0.994. The LP coefficients are converted into LSFs for efficient quantisation, as discussed in Section 5.4.

The gain is estimated based on the root-mean-squared (RMS) value of the pitch cycle waveform for voiced signals. On the other hand, for unvoiced signals, the RMS is calculated with a fixed number of samples, e.g. 80.

The spectral magnitudes are estimated for each harmonic frequency of voiced speech. The magnitudes for unvoiced speech are estimated for a fixed frequency interval of 100
For an input spectrum $S(k)$, $0 < k < K$, the estimated spectral magnitude $A(h)$ for the $h$th harmonic is given by

$$A(h) = \sqrt{\frac{\sum_{k=1}^{u_h} |S(k)|^2 w_h(k - h\omega_0)}{\sum_{k=1}^{u_h} w_h(k - h\omega_0)}}, \quad \text{for } 1 \leq h \leq H,$$

(6.6)

where the fundamental frequency $\omega_0$ and the total number of harmonics $H$ are calculated from the pitch period $T_0$ as $\omega_0 = 2\pi/T_0$ and $H = \lfloor T_0/T \rfloor$, respectively. The lower and upper frequency bounds, $l_h$ and $w_h$, of the $h$th harmonic are given

$$l_h = \left\lfloor \frac{K(h - 0.5)\omega_0}{2\pi} + 0.5 \right\rfloor$$

(6.7)

and

$$w_h = \left\lfloor \frac{K(h + 0.5)\omega_0}{2\pi} - 0.5 \right\rfloor,$$

(6.8)

respectively. In (6.6), the harmonic magnitude is calculated based on the weighted expectation of $|S(k)|^2$ depending on the voicing decision for each harmonic. In specific, the weights for each voiced and unvoiced harmonics are defined with the sync and flat functions, respectively, as

$$w_h(\omega) = \begin{cases} \sin(\alpha\omega) / \alpha, & \text{if } h\text{th harmonic is voiced,} \\ 1, & \text{otherwise} \end{cases}$$

(6.9)

where $\alpha = 2\pi T_0/K$.

### 6.6 Quantisation of Model Parameters

Efficient quantisation with minimum number of bits while retaining the perceptual speech-quality is crucial in designing speech codecs.

The pitch is quantised in terms of the number of samples in the time domain. In consideration of the frequency distortion, the pitch is quantised in the logarithmic scale using 7 bits for the second subframe. On the other hand, the pitch for the first subframe is quantised using the differential coding method. In other words, the pitch difference between the second and first subframes is quantised using 5 bits.
The voicing level for each subframe is quantised using 3 bits, corresponding to the resolution of 7 frequency sub-bands.

The LSFs are quantised into 24 bit/frame by MSVQ combined with the switched predictive (SP) scheme as described in Section 5.4.5. The predictive scheme is introduced in order to exploit the interframe redundancy in consecutive LSF vectors. Auto-regressive (AR) prediction is considered because it can remove the interframe correlation more effectively compared to the moving average (MA) scheme. It may, however, cause propagating the effect of channel errors to the consecutive frames. To circumvent this problem, a switching scheme between the AR predictive and memoryless quantisers is finally employed. Moreover, the switching scheme has the advantage of selecting an optimum one for various kinds of speech signals. In other words, the switching scheme could contribute to further reducing the spectral distortion by a proper selection of either AR predictive or memoryless scheme depending on the characteristics of speech signals.

MA prediction is applied prior to gain quantisation, considering that the gain of the speech signal changes slowly from frame to frame. The residual term of the predictor is quantised in the logarithmic scale with 5 bits/subframe.

Normalised spectral magnitudes, the frequency response of the LP-residual signal, is quantised into 14 bit/subframe by SP-MVQ, proposed in Section 5.7.

Bit allocations for the 4 kbit/s speech and 1.7 kbit/s CNI (see Section 6.9) frames are shown in Table 6.1 and 6.2, respectively.

### 6.7 Packetisation and Frame Erasure Handling

The quantised model parameters are packed into a bit stream using packetisation to transmit through the telecommunication channels. In the case of DTX, it generates three kinds of packets: speech, CNI, and silence. Each packet mode can be detected at the decoder by inspecting the header of received packets. The payload sizes of 80, 34, and 0 indicate speech, CNI, and silence frames, respectively. The packet structure is shown in Table 6.3.
6.7. Packetisation and Frame Erasure Handling

Table 6.1: Bits allocation of the eSB-LPC vocoder for voice-active frames.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits/frame (subframe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP coefficients</td>
<td>24</td>
</tr>
<tr>
<td>Pitch</td>
<td>12 (5 + 7)</td>
</tr>
<tr>
<td>Voicing</td>
<td>6 (3 + 3)</td>
</tr>
<tr>
<td>RMS Gain</td>
<td>10 (5 + 5)</td>
</tr>
<tr>
<td>Spectral magnitude</td>
<td>28 (14 + 14)</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 6.2: Bits allocation of the eSB-LPC vocoder for CNI frames.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits/frame (subframe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP coefficients</td>
<td>24</td>
</tr>
<tr>
<td>RMS-Gain</td>
<td>10 (5 + 5)</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 6.3: Packet structure of the eSB-LPC vocoder.

<table>
<thead>
<tr>
<th>Contents</th>
<th>Allocated number of bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame Sync (0x6B21)</td>
<td>16</td>
</tr>
<tr>
<td>Size</td>
<td>16</td>
</tr>
<tr>
<td>Payload</td>
<td>80, 34, or 0</td>
</tr>
</tbody>
</table>
Safe transmission of a frame can be detected by checking the frame sync. A frame is classified into erasure if the field for the sync word is not 0x6B21. The frame repetition technique, reusing the most recent safe packet, is employed for resilience against channel errors in the case of frame erasure.

### 6.8 Speech Synthesis

Speech is synthesised with pitch cycle waveform (PCW)-based parameter interpolation. Firstly, intermediate PCWs for the current subframe are generated by interpolating the quantised model parameters of the last and current subframes. The excitation signal $e_i(n), 0 \leq n < T_{b,i}$, for the $i$th PCW is produced as

$$
e_i(n + n_i) = \sum_{h=1}^{V_c} A_{a,i}(h) \cos \{h\omega_{0,i}(n - n_i)\} + \sum_{h=V_c + 1}^{H} A_{e,i}(h) \cos \{h\omega_{0,i}(n - n_i) + U[-\pi, \pi]\}$$

(6.10)

where $\omega_{0,i} = 2\pi/T_{b,i}$, and $U[-\pi, \pi]$ denotes a random number with uniform distribution between $-\pi$ and $\pi$. The start position $n_i$ for the $i$th PCW is given by

$$n_i = n_0 + \sum_{j=0}^{i-1} T_{b,j}$$

(6.11)

where $n_0$ is the start position corresponding to the last position of the previous subframe. The interpolated pitch $T_{b,i}$ for the $i$th PCW is calculated as

$$T_{b,i} = \alpha_i T_0^{(t-1)} + (1 - \alpha_i) T_0^{(t)}$$

(6.12)

where $T_0^{(t)}$ is the quantised pitch of the $t$th subframe. The interpolation factor $\alpha_i$ is defined

$$\alpha_i = \frac{\hat{G}^{(t)} N_i}{\hat{G}^{(t-1)} (N - N_i) + \hat{G}^{(t)} N_i}$$

(6.13)

where $N$ is the subframe size, $\hat{G}^{(t)}$ is the quantised gains, and $N_i$ is the PCW position defined

$$N_i = n_i + 0.25(T_0^{(t-1)} + T_0^{(t)}).$$

(6.14)
The start position \( n_0^{(t+1)} \) for the next subframe is updated as

\[
\begin{align*}
n_0^{(t+1)} &= (n_j^{(t)} + T_{0,j}) \% N \\
\end{align*}
\]

where \( \% \) is the modulo operator, and \( I \) is the total number of PCWs. The voicing cutoff index \( V_c \) is given by

\[
\begin{align*}
V_c &= \max\{\hat{V}_c^{(t-1)}, \hat{V}_c^{(t)}\}. \\
\end{align*}
\]

The interpolated magnitude \( A_{e_i}(h) \) for the \( h \)th harmonic is computed as

\[
\begin{align*}
A_{e_i}(h) &= \begin{cases} 
\alpha_i\hat{A}_e^{(t-1)}(h) + (1 - \alpha_i)\hat{A}_e^{(t)}(h), & \text{if } \hat{V}_c^{(t-1)}(h) = \hat{V}_c^{(t)}(h), \\
\hat{A}_e^{(t-1)}(h), & \text{if } \hat{V}_c^{(t-1)}(h) = 1 & \hat{V}_c^{(t)}(h) = 0, \\
\hat{A}_e^{(t)}(h), & \text{if } \hat{V}_c^{(t-1)}(h) = 0 & \hat{V}_c^{(t)}(h) = 1,
\end{cases} \\
\end{align*}
\]

where \( \hat{V}_c^{(t)}(h) \) is the voicing information for the \( h \)th harmonic; 1 and 0 in the voicing comparison denote voiced and unvoiced, respectively. The LP-coefficient for the \( i \)th PCW is interpolated in the same way obtaining the interpolated pitch. Actually, the interpolation is performed using LSFs instead of LP coefficients, followed by the conversion from LSFs to LP coefficients. The conversion is conducted by comparing the impulse responses of the LSP and the LP synthesis filters [110]. Normalised speech signal \( \tilde{s}_i(n) \) is reconstructed by exciting the LP synthesis filter \( h_i(n) \) with the signal \( e_i(n) \) in (6.10), as

\[
\tilde{s}_i(n) = e_i(n) \ast h_i(n) \\
\]

where * is the convolution operator. In calculation of \( \tilde{s}_i(n) \), the required memory for \( e_i(n) \), \( n < 0 \), can be obtained from \( e_{i-1}(n) \) or the excitation signal of the last subframe. Speech signals \( s_i(n) \) for the \( i \)th PCW is produced by compensating for the gain as

\[
\begin{align*}
\tilde{s}_i(n) &= \sqrt{\frac{T_{0,i}}{\sum_{n=0}^{T_{0,i}-1} \tilde{s}_i^2(n)}} G_i \tilde{s}_i(n) \\
\end{align*}
\]

where \( G_i \) is the interpolated gain based on the relative position of the PCW in the subframe. Concatenation of each PCW in (6.19) forms the final speech signal.
6.9 Comfort Noise Generation and Insertion

Comfort noise generation (CNG) enables to produce natural synthesised speech for silence frames by hiding discontinuities between speech and silence frames. VAD combined with CNG enables to deploy the silence compression scheme in communication systems without speech quality degradation while achieving reduction in the average bit-rate. The technique employed for CNG is the replication of the last silence frame information. For consecutive silence frames, the speech decoder takes the frame information from the most recent safe packet. Note that in CNG, the spectral phases for speech synthesis are set to random numbers of $U[-\pi, \pi]$.

Background noise may exhibit the change in its characteristics. Thus, comfort noise insertion (CNI) is introduced every certain number of consecutive silence frames to produce more natural sound by reflecting a portion of current frame information. The information for CNI includes minimal characteristics over the background noise such as the gain and LP coefficients. Note that, in producing CNI frames, the phase and spectral magnitudes are set to random and flat values, respectively.

6.10 Subjective Speech-Quality Evaluation

A subjective test is conducted to evaluate speech quality of the proposed eSB-LPC vocoder. For the listening test, four sentences uttered by two-male and two-female speakers are used, filtered by m-IRS, mixed with noise signals, and then processed by speech coders. Each speech material for presentation to listeners has the length of 8 sec, composed of two 4-sec sentences. There is 1 ~ 2 sec pause between the talk spurts. The noise is generated by vehicle and babble sources with 10 dB SNR. The speech coders are composed of eSB-LPC, EVRC [28], G.729 [12], and G.723.1 [10] of which the maximum bit-rates are 4, 8.55, 8, and 5.3 kbit/s, respectively. The eSB-LPC vocoder is further classified into four operating modes depending on the switches for both VAD and noise reduction. In the case of EVRC, both variable-rate and noise reduction modes are set to being enabled. The listener group is composed of 12 people. Each listener is asked to mark his/her mean opinion score (MOS) for each randomly
Table 6.4: Mean opinion score.

<table>
<thead>
<tr>
<th>MOS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Very good</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
</tr>
</tbody>
</table>

Presented sentence-pair of the processed speech material. The MOS ranges between 1 and 5, as shown in Table 6.4.

The speech quality test results are shown in Table 6.5. The eSB-LPC without noise reduction results in performances inferior to both G.729 and EVRC, but comparable to G.723.1. However, when the noise reduction mode is enabled, its performance is comparable to EVRC for noisy speech signals. The average bit-rates of both eSB-LPC and EVRC are given in Table 6.6. It shows that, for noisy speech signals, eSB-LPC operates lower than half the bit-rate of EVRC while producing speech with comparable quality.

### 6.11 Summary

This chapter has proposed an eSB-LPC vocoder, featuring VAD, noise reduction, improved pitch estimation, and efficient spectral magnitude quantisation. It operates at 4, 1.7, and 0 kbit/s for speech, CNI and silence frames, respectively. The eSB-LPC vocoder conducts speech enhancement for quality improvement under noise environment. The speech quality of the proposed speech coder has been compared with those of three kinds of CELP coders, including EVRC, G.729, and G.723.1. The eSB-LPC vocoder without noise reduction results in performances inferior to G.729, but comparable to G.723.1. However, with noise reduction, the eSB-LPC vocoder produces speech qualities comparable to EVRC for noisy speech signals while operating at lower
Table 6.5: Subjective speech-quality test results in term of the MOS.

<table>
<thead>
<tr>
<th>Speech codec</th>
<th>VAD</th>
<th>NR</th>
<th>Clean</th>
<th>Vehicle 10 dB SNR</th>
<th>Babble 10 dB SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>All</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>eSB-LPC</td>
<td>on</td>
<td>on</td>
<td>3.81</td>
<td>3.79</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>off</td>
<td>on</td>
<td>3.88</td>
<td>3.92</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>on</td>
<td>off</td>
<td>3.73</td>
<td>3.75</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td>off</td>
<td>off</td>
<td>3.77</td>
<td>3.88</td>
<td>3.67</td>
</tr>
<tr>
<td>EVRC</td>
<td>on</td>
<td>on</td>
<td>4.35</td>
<td>4.38</td>
<td>4.33</td>
</tr>
<tr>
<td>G.723.1</td>
<td>off</td>
<td>off</td>
<td>4.13</td>
<td>4.17</td>
<td>4.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.77</td>
<td>3.71</td>
<td>3.83</td>
</tr>
</tbody>
</table>

(NR: noise reduction, M: male, F: female)

Table 6.6: Comparison of the average bit-rates between the EVRC and eSB-LPC speech coders with the VAD mode on.

<table>
<thead>
<tr>
<th>Speech codec</th>
<th>Average bit-rate (kbit/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clean</td>
</tr>
<tr>
<td>eSB-LPC</td>
<td>2.45</td>
</tr>
<tr>
<td>EVRC</td>
<td>6.02</td>
</tr>
</tbody>
</table>
than half the average-bit-rate of EVRC.
Chapter 7

Conclusions and Future Research

7.1 Conclusions

This thesis has proposed advanced techniques for VAD, speech enhancement, and speech model parameters estimation and quantisation. Based on the proposed techniques, an eSR-LPC vocoder has been developed.

VAD together with CNG presents considerable reduction in required bit-rates for voice communications. LR-based VAD exploits a statistical approach for the detection of speech frames, while traditional VAD methods tackle the detection in heuristic ways. However, it has been found that the LR-based VAD method may malfunction at speech offset regions. Thus, this thesis has identified the reason for the unwanted phenomenon, and then has proposed SLR-based VAD to alleviate the problem. Objective quality test results have shown that the proposed method gives performances significantly better than LR-based VAD and comparable to ETSI AMR VAD option 2.

Background noise could cause considerable degradation in the voice communication quality. Suppression of the noise level from input speech can give more pleasant and intelligible voice communications. Traditionally, noise reduction techniques have been studied mainly for enhanced spectral estimation over given noise statistics. The noise statistics are typically adapted during voice-inactive regions by means of VAD. However, VAD errors may cause considerable mis-estimation of the noise statistics. Hence, this
thesis has proposed a mixed-decision based noise adaptation method for robustness against VAD errors. The proposed method is designed on the basis of an integration of soft- and hard-decision techniques. The usefulness of the proposed methods has been proven through objective speech-quality measurement in terms of both SEGSNR improvement and ISD.

Speech compression presents main reduction in the channel bandwidth for voice communications. One of the most crucial parameters in most low bit-rate speech coders is the pitch. Depending on its estimation accuracy, the speech quality could be affected substantially. Hence, this thesis has proposed a pitch estimation method based on SA. By combining SA with other pitch estimation metrics, such as TA or SS, it has been shown that the proposed methods enable to reduce a considerable number of pitch estimation errors over various noise sources with levels.

The sinusoidal speech model is very well-known for low bit-rate coding. As one of the main parameters of sinusoidal speech coders, the spectral magnitudes of LP-residual signals play an important role in the quality of synthesised speech. Hence, this thesis has proposed two methods, PMBVQ and SP-MVQ, for efficient quantisation of the spectral magnitudes. Both of them are designed based on mel-scale dimension conversion to alleviate the dimension variability of spectral vectors. Moreover, both methods employ predictive coding schemes in order to exploit the statistical redundancy between consecutive frames. Through a subjective speech-quality test, it has been shown that SP-MVQ contributes considerably to improving the speech quality of a sinusoidal speech coder.

Finally, this thesis has proposed an eSB-LPC vocoder incorporating the above proposed techniques. It operates at the bit rates of 4, 1.7, and 0 kbit/s for speech, CNI, and silence frames, respectively. The subjective speech-quality of the proposed vocoder has been evaluated, and then compared with those of a few CELP coders, as EVRC, G.729, and G.723.1. The eSB-LPC vocoder without noise reduction results in performances comparable to G.723.1 but inferior to G.729. However, with noise reduction, the eSB-LPC vocoder gives speech quality comparable to EVRC for noisy speech signals whilst operating at lower than a half of the average bit-rate of EVRC.
7.2 Future Research

- **VAD and speech enhancement against non-stationary noise signals:** The noise signal assumed in this thesis is restricted to stationary or slow-varying ones. However, in real environments, it could be non-stationary noise in which the statistics change rapidly from frame to frame. The VAD and speech enhancement techniques introduced in Chapter 3 and 4 are, however, not feasible for non-stationary noise signals. Thus, a novel spectral estimation technique together with noise adaptation against non-stationary noise signals would be a challenging subject.

- **Phase modelling for transitional speech signals:** The eSB-LPC vocoder gives speech quality comparable to EVRC for noisy speech. It is due to the high performance of the speech enhancement module, adopted as a preprocessor of the speech encoder. However, for clean speech signals, the eSB-LPC vocoder gives performances worse than CELP-based coders operating circa 8 kbit/s, such as EVRC and G.729. The reason is that the zero-and-random phase model of sinusoidal speech coders does not seem to properly cope with transitional regions of speech. Consequently, it would give degradation in overall speech-quality because of inconsistent sounds produced at the transitional regions. The transitional region includes the transitions between unvoiced and voiced speeches, plosive sounds such as /p/ and /b/, and plosive fricative sounds such as /t/ and /d/. In order to alleviate this problem, it could be a good solution if the phase of the signal is included in modelling the speech. Proper handling of transitional speech seems to require a more complicated phase model for a higher speech-quality. However, it is not easy to model and quantise the phase terms of speech signals. A general model together with efficient quantisation of the phase is expected to present a considerable advance in sinusoidal model-based speech coding.

- **Network-driven variable bit-rate speech coding:** Network-driven variable rate control could be a crucial concern for future voice coders, considering that communication systems move toward packet-based networks on wideband CDMA. It is a well-known fact that sinusoidal speech coders produce a reasonably good
sound at the bit rates around 2 kbit/s, compared to other speech coding methods. A sinusoidal codec with a solution for the transitional speech could cover a range of bit rates about 1 ~ 8 kbit/s without serious quality-variance, while CELP coders exhibits relatively severe degradation of the quality for a decrease of the bit rate, especially below 4 kbit/s. Hence, it seems that sinusoidal model-based approaches are desirable for developing variable bit-rate speech coders over packet and wireless networks liable to packet congestion, high radio interference, or both.
# Appendix A

## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AbS</td>
<td>analysis-by-synthesis</td>
</tr>
<tr>
<td>ADPCM</td>
<td>adaptive differential pulse code modulation</td>
</tr>
<tr>
<td>AMBE</td>
<td>advanced multi-band excitation</td>
</tr>
<tr>
<td>AMR</td>
<td>adaptive multi-rate</td>
</tr>
<tr>
<td>AMR1</td>
<td>AMR VAD option 1</td>
</tr>
<tr>
<td>AMR2</td>
<td>AMR VAD option 2</td>
</tr>
<tr>
<td>AR</td>
<td>auto-regressive</td>
</tr>
<tr>
<td>CDMA</td>
<td>code division multiple access</td>
</tr>
<tr>
<td>CELP</td>
<td>code-excited linear prediction</td>
</tr>
<tr>
<td>CNG</td>
<td>comfort noise generation</td>
</tr>
<tr>
<td>CNI</td>
<td>comfort noise insertion</td>
</tr>
<tr>
<td>DC</td>
<td>direct current</td>
</tr>
<tr>
<td>DCT</td>
<td>discrete cosine transform</td>
</tr>
<tr>
<td>DD</td>
<td>decision-directed</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>DoD</td>
<td>department of defense</td>
</tr>
<tr>
<td>DSVD</td>
<td>digital simultaneous voice and data</td>
</tr>
<tr>
<td>DTX</td>
<td>discontinuous transmission</td>
</tr>
</tbody>
</table>
EDGE  enhanced data rates for GSM evolution
EFR   enhanced full rate
EIA   electronic industries association
eSB-LPC  enhanced split-band linear predictive coding
ETSI  European telecommunications standards institute
EVRC  enhanced variable bit-rate coder
FFT   fast Fourier transform
FR    full-rate
FS    federal standard
GBSS  generalised Berouti's spectral subtraction
GPRS  general packet radio service
GSM   global system for mobile communications
GSS   generalised spectral subtraction
G.723.1A  G.723.1 Annex A
G.729B  G.729 Annex B
HAS   human auditory system
HD    hard decision
HO    hangover
HPF   high pass filter
HR    half-rate
IMBE  improved multi-band excitation
INMARSAT  international maritime satellite corporation
IS    interim standard
ISD   Itakura-Saito distortion
ISDN  integrated services digital network
ITU-T international telecommunication union - telecommunication
KLT   Karhunen-Loève transform
LP    linear predictive
LPC   linear predictive coding
LR    likelihood ratio
LSA   log spectral amplitude
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSF</td>
<td>line spectrum frequency</td>
</tr>
<tr>
<td>LSP</td>
<td>line spectrum pair</td>
</tr>
<tr>
<td>LSVQ</td>
<td>linked split vector quantisation</td>
</tr>
<tr>
<td>LTP</td>
<td>long-term prediction</td>
</tr>
<tr>
<td>MA</td>
<td>moving average</td>
</tr>
<tr>
<td>MBCB</td>
<td>mel-scale codebook</td>
</tr>
<tr>
<td>MBE</td>
<td>multi-band excitation</td>
</tr>
<tr>
<td>MBE-LPC</td>
<td>multi-band excited linear predictive coding</td>
</tr>
<tr>
<td>MBVQ</td>
<td>mel-scale binary vector quantisation</td>
</tr>
<tr>
<td>MD</td>
<td>mixed decision</td>
</tr>
<tr>
<td>MELP</td>
<td>mixed-excitation linear prediction</td>
</tr>
<tr>
<td>m-IRS</td>
<td>modified input response system</td>
</tr>
<tr>
<td>ML</td>
<td>maximum likelihood</td>
</tr>
<tr>
<td>MMSE</td>
<td>minimum mean square error</td>
</tr>
<tr>
<td>MOS</td>
<td>mean opinion score</td>
</tr>
<tr>
<td>MPE</td>
<td>multi-pulse excitation</td>
</tr>
<tr>
<td>MSVQ</td>
<td>multi-stage vector quantisation</td>
</tr>
<tr>
<td>PCM</td>
<td>pulse code modulation</td>
</tr>
<tr>
<td>PCW</td>
<td>pitch cycle waveform</td>
</tr>
<tr>
<td>PDA</td>
<td>pitch determination algorithm</td>
</tr>
<tr>
<td>PDF</td>
<td>probability density function</td>
</tr>
<tr>
<td>PMBVQ</td>
<td>predictive mel-scale binary vector quantisation</td>
</tr>
<tr>
<td>P-MVQ</td>
<td>predictive - mel-scale vector quantisation</td>
</tr>
<tr>
<td>PSAP</td>
<td>a priori speech absence probability</td>
</tr>
<tr>
<td>PSS</td>
<td>power spectral subtraction</td>
</tr>
<tr>
<td>PSTN</td>
<td>public switched telephone network</td>
</tr>
<tr>
<td>PWF</td>
<td>perceptual weighting filter</td>
</tr>
<tr>
<td>QCELP</td>
<td>Qualcomm code-excited linear prediction</td>
</tr>
<tr>
<td>RDA</td>
<td>rate determination algorithm</td>
</tr>
<tr>
<td>RMS</td>
<td>root mean square</td>
</tr>
<tr>
<td>RPE-LPC</td>
<td>regular pulse-excited linear predictive coding</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>SA</td>
<td>spectral autocorrelation</td>
</tr>
<tr>
<td>SAP</td>
<td>speech absence probability</td>
</tr>
<tr>
<td>SAS</td>
<td>strongly active speech</td>
</tr>
<tr>
<td>SB-ADPCM</td>
<td>sub-band adaptive differential pulse code modulaltion</td>
</tr>
<tr>
<td>SB-LPC</td>
<td>split-band linear predictive coding</td>
</tr>
<tr>
<td>SD</td>
<td>soft decision</td>
</tr>
<tr>
<td>SEGSNR</td>
<td>segmental signal to noise ratio</td>
</tr>
<tr>
<td>SLR</td>
<td>smoothed likelihood ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>signal to noise ratio</td>
</tr>
<tr>
<td>SP</td>
<td>switched-predictive</td>
</tr>
<tr>
<td>SP-MVQ</td>
<td>switched-predictive mel-scale vector quantisation</td>
</tr>
<tr>
<td>SS</td>
<td>spectral synthesis</td>
</tr>
<tr>
<td>SS-SA</td>
<td>spectral synthesis - spectral autocorrelation</td>
</tr>
<tr>
<td>STA</td>
<td>spectro-temporal autocorrelation</td>
</tr>
<tr>
<td>STC</td>
<td>sinusoidal transform coder</td>
</tr>
<tr>
<td>STSA</td>
<td>short-time spectral amplitude</td>
</tr>
<tr>
<td>TA</td>
<td>temporal autocorrelation</td>
</tr>
<tr>
<td>TDMA</td>
<td>time division multiple access</td>
</tr>
<tr>
<td>TIA</td>
<td>telecommunication industries association</td>
</tr>
<tr>
<td>UMTS</td>
<td>universal mobile telecommunication service</td>
</tr>
<tr>
<td>UV</td>
<td>unvoiced</td>
</tr>
<tr>
<td>VAD</td>
<td>voice activity detection</td>
</tr>
<tr>
<td>VBR</td>
<td>variable bit-rate</td>
</tr>
<tr>
<td>VoIP</td>
<td>voice over internet protocol</td>
</tr>
<tr>
<td>VQ</td>
<td>vector quantisation</td>
</tr>
<tr>
<td>WF</td>
<td>Wiener filtering</td>
</tr>
<tr>
<td>WI</td>
<td>waveform interpolation</td>
</tr>
<tr>
<td>WT</td>
<td>wavelet transform</td>
</tr>
<tr>
<td>WTA</td>
<td>weighted temporal autocorrelation</td>
</tr>
<tr>
<td>WSNR</td>
<td>weighted signal to noise ratio</td>
</tr>
<tr>
<td>WSS</td>
<td>weighted spectral synthesis</td>
</tr>
<tr>
<td>ZC</td>
<td>zero crossing</td>
</tr>
</tbody>
</table>
Appendix B

List of Publications


7. S. Villette, Y. D. Cho, A. Kondoz, "Efficient Parameter Quantisation for 2.4/1.2 kb/s Split-band LPC Coding," *IEEE Workshop on Speech Coding*, Delavan, WI,
USA, pp. 32-4, Sep. 2000


Bibliography


[19] ETSI, "Digital cellular telecommunications system (phase 2+); full rate speech transcoding," *GSM 06.10 (ETS 300 530-2)*.

[20] ETSI, "Digital cellular telecommunications system (phase 2+); Half rate speech; Half rate speech transcoding," *GSM 06.20 v5.1.0 (draft ETSI ETS 300 969)*, Dec 1997.
[21] ETSI, “Digital cellular telecommunications system (phase 2); Enhanced full rate (EFR) speech transcoding,” *GSM 06.60 v4.1.0 (ETSI 301 245)*, June 1998.

[22] ETSI, “Digital cellular telecommunications system (phase 2+); Adaptive multi-rate (AMR) speech transcoding,” *GSM 06.90 v7.2.0 (draft ETSI EN 301 704)*, 1998.

[23] ETSI, “Digital cellular telecommunications system (phase 2+); Voice activity detector (VAD) for full rate speech traffic channels,” *GSM 06.32 (ETSI EN 300 965 v7.0.1)*, 1998.

[24] ETSI, “Digital cellular telecommunications system (phase 2+); Voice activity detector (VAD) for full rate speech traffic channels,” *GSM 06.42 (draft ETSI EN 300 973 v8.0.0)*, 1999.


[71] I. Atkinson, Advanced linear predictive speech compression at 3.0 kbit/s and below.


