

GLOTTAL-SYNCHRONOUS SPEECH PROCESSING

by
MARK R. P. THOMAS
M.Eng (Hons)

A Thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy of University of London and
Diploma of Imperial College

Communications and Signal Processing Research Group
Department of Electrical Engineering
Imperial College London
2010

Abstract

Glottal-synchronous speech processing is a field of speech science where the pseudoperiodicity of voiced speech is exploited. Traditionally, speech processing involves segmenting and processing short speech frames of predefined length; this may fail to exploit the inherent periodic structure of voiced speech which glottal-synchronous speech frames have the potential to harness. Glottal-synchronous frames are often derived from the glottal closure instants (GCIs) and glottal opening instants (GOIs).

The SIGMA algorithm was developed for the detection of GCIs and GOIs from the Electroglottograph signal with a measured accuracy of up to 99.59%. For GCI and GOI detection from speech signals, the YAGA algorithm provides a measured accuracy of up to 99.84%. Multichannel speech-based approaches are shown to be more robust to reverberation than single-channel algorithms.

The GCIs are applied to real-world applications including speech dereverberation, where SNR is improved by up to 5 dB, and to prosodic manipulation where the importance of voicing detection in glottal-synchronous algorithms is demonstrated by subjective testing. The GCIs are further exploited in a new area of data-driven speech modelling, providing new insights into speech production and a set of tools to aid deployment into real-world applications. The technique is shown to be applicable in areas of speech coding, identification and artificial bandwidth extension of telephone speech.

In loving memory of Pauline Rose Gunn

11/12/1952 – 08/10/2009

Acknowledgments

Foremost, I would like to thank my supervisor, Patrick Naylor, for encouraging me to undertake this PhD, for guiding me as a researcher and for supporting the latest idea. I also owe a great debt to my friends and colleagues Nikolay Gaubitch and Jon Gudnason for their collaborative research and guidance with the finer details of a PhD.

I would additionally like to thank my friends Beth, David, Emanuel, Jesse, Jimi and Paul for the good times in and out of the office. I extend my gratitude back home to my parents, Richard and Janet, my girlfriend Kate and my friend Palo, for supporting my work, for occasionally having to feign interest and enduring my temperament during the writing of this thesis.

And finally thank you to Grandma, Nanny, Pauline and Max, whom I have lost these past three years. I hope you're safe and I've made you proud.

Contents

Abstract	2
Dedication	3
Acknowledgments	4
Contents	5
List of Figures	10
List of Tables	13
Statement of Originality	14
Chapter 1. Introduction	27
1.1 Context of Work	27
1.1.1 Voice Modelling	27
1.1.2 Application Context	28
1.2 Research Statement and Thesis Structure	29
1.2.1 Research Statement	29
1.2.2 Thesis Structure	30
Chapter 2. Review	31
2.1 Introduction	31
2.2 Physiology of Voice Production	31
2.2.1 Physiology of the Vocal Tract	33
2.3 Models of Speech Production	34
2.3.1 The Linear Source-Filter Model	35
2.4 Modelling the Vocal Tract Transfer Function	37
2.5 Glottal Waveform Models	41
2.5.1 Curve Fitting Models	42
2.5.2 Physical Modelling	43

2.5.3	Error-Minimizing Models	44
2.6	Estimating Vocal Tract Transfer Function and Glottal Waveforms	45
2.6.1	Linear Predictive Coding (LPC)	45
2.6.2	Simultaneous Estimation of Vocal Tract and Glottal Excitation	50
2.6.3	Homomorphic Processing	51
2.7	Measuring Glottal Waveforms (Invasive Methods)	52
2.7.1	EKG	53
2.7.2	Pneumotachograph	54
2.7.3	Laryngoscopy	54
2.8	GCI/GOI Detection	56
2.8.1	Noninvasive Methods	56
2.8.2	Invasive Methods	57
2.8.3	Voiced/Unvoiced/Silence Detection	57
2.8.4	Pitch Detection	59
2.9	Application Context	60
2.9.1	Dereverberation	60
2.9.2	Time Scale Modification	63
2.9.3	Artificial Bandwidth Extension	64
2.10	Speech Quality Measurements	65
2.10.1	Objective Measures	65
2.10.2	Subjective Measures	67
2.11	Training and Test Corpora	68
2.12	Chapter Summary	69
Chapter 3.	Glottal Activity Detection from Electroglottograph Signals	70
3.1	Introduction	70
3.2	Problem Discussion	71
3.2.1	Defining the GCI	72
3.2.2	Defining the GOI	73
3.2.3	Errors in Existing Approaches	73
3.3	Glottal Activity Detection with the SIGMA Algorithm	79
3.3.1	Multiscale Analysis	79
3.3.2	Group Delay Function	82
3.3.3	Candidate Selection	84
3.3.4	Swallowing	86
3.3.5	GOI Post-Filtering	86
3.4	Results and Discussion	87
3.4.1	Experiment 1: Evaluation with APLAWD and SAM	88

3.4.2	Experiment 2: Variation in Group Delay Window Size	91
3.4.3	Experiment 3: Comparison with Cubic Spline Wavelet	93
3.5	Chapter Summary	93
Chapter 4.	Glottal Activity Detection from Speech Signals	95
4.1	Introduction	95
4.2	Problem Discussion and Existing Approaches	96
4.2.1	Preprocessing	96
4.2.2	Event Detection	97
4.2.3	Postprocessing	98
4.2.4	Failure Modes	98
4.3	The DYPSA Algorithm	99
4.3.1	Preprocessor	99
4.3.2	Event Detection	99
4.3.3	Postprocessor	100
4.4	The YAGA Algorithm	103
4.4.1	Preprocessor	104
4.4.2	Event Detection	105
4.4.3	Postprocessing	105
4.4.4	GOI Detection	108
4.4.5	Evaluation	112
4.5	GCI Detection from Reverberant Speech	117
4.5.1	Problem Formulation	117
4.5.2	DYPSA at the output of a beamformer	118
4.5.3	Multichannel DYPSA	119
4.5.4	Evaluation	121
4.6	Chapter Summary	125
Chapter 5.	Data-Driven Voice Source Modelling	126
5.1	Introduction	126
5.2	Preliminary Processing	128
5.2.1	Concatenation of Voice Source Frames for Resynthesis	129
5.3	Feature Modelling	131
5.3.1	Model Complexity	132
5.3.2	Prototype Voice Source Signals	133
5.3.3	Analysis-Synthesis	133
5.4	Transform Modelling	135
5.4.1	Principle Component Analysis	135
5.4.2	GMM Analysis on PCA Spectra	138

5.4.3	Alternative Orthogonal Transforms	139
5.5	Experimentation	141
5.5.1	An Enhanced LPC Preemphasis Filter	141
5.5.2	Analysis-Synthesis	144
5.5.3	GMM Analysis	144
5.5.4	Speaker Dependence	145
5.6	Chapter Summary	146
Chapter 6.	Applications of Glottal-Synchronous Processing	148
6.1	Introduction	148
6.2	Application 1: A Practical Multichannel Dereverberation Algorithm	149
6.2.1	Problem Formulation	150
6.2.2	Proposed Approach	151
6.2.3	Evaluation	154
6.3	Application 2: Time Scale Modification of Speech	156
6.3.1	Problem Formulation	157
6.3.2	Proposed Approach	159
6.3.3	Evaluation	161
6.4	Application 3: Artificial Bandwidth Extension of Telephone Speech	164
6.4.1	Existing Voice Source ABWE Techniques	166
6.4.2	Proposed Model-Based Source Extension	167
6.4.3	Proposed Temporal & Spectral Envelope Estimation	169
6.4.4	Evaluation	171
6.5	Chapter Summary	173
Chapter 7.	Conclusions	175
7.1	Résumé	175
7.2	Further Work	178
7.3	Publications	180
7.3.1	Conferences	180
7.3.2	Journals	181
7.3.3	Contributions to Books	181
Bibliography		182
Appendix A.	SIGMA Group Delay Window for Male and Female Speech	202
A.1	Varying Group Delay Window for Male Speech	203
A.2	Varying Group Delay Window for Female Speech	204
Appendix B.	GMM-Derived Voice Source Prototypes	205

B.1 Feature Modelling with MFCCs	206
B.2 Transform Modelling with PCA	207

List of Figures

2.1	Cross section schematics of the larynx and a reed organ pipe	32
2.2	The Glottal Cycle	33
2.3	Schematic diagram of human speech physiology	34
2.4	Block diagram of human speech production	35
2.5	The source-filter model of speech	36
2.6	Discretized cross-section of the vocal tract	37
2.7	Pole plot and frequency response of vocal tract	40
2.8	LF model of the glottal waveform	41
2.9	Speech signal and estimated glottal waveforms	49
2.10	Speech signal and EGG	53
2.11	High-speed laryngoscopy	55
2.12	Process of reverberation	60
2.13	Effects of reverberation on LPC residuals	61
2.14	Speech spectra for unvoiced and voiced speech	64
3.1	Speech, EGG and EGG time derivative	72
3.2	Speech, EGG and EGG time derivative for breathy offset	75
3.3	Speech, EGG and EGG derivative for breathy voice	76
3.4	Time-stretching with errors due to incorrect GCIs	77
3.5	Discrete and Stationary Wavelet Transforms	80
3.6	SWT analysis and detail filters	81
3.7	EGG, multiscale product and group delay fn. with overlaid candidates . .	82
3.8	Distribution of EGG GCI feature vectors	85
3.9	SIGMA system diagram	87
3.10	GCI and GOI evaluation strategy	89
3.11	Effect of varied group delay window length on evaluation measures	92
4.1	Variation of group delay window length	106
4.2	Voicing detector using waveform similarity measure	107

4.3	Definitions of glottal opening	109
4.4	Voice source for normal modal and breathy voice	110
4.5	GOI candidates as CQ with best path overlaid	111
4.6	Voice source, group delay fn. and multiscale product w/overlaid candidates	112
4.7	Proposed GCI/GOI detection algorithm system diagram	113
4.8	Error histograms for GCI and GOI detection	114
4.9	Hit rate and accuracy with additive babble noise	116
4.10	Hit rate and accuracy with additive factory noise	116
4.11	Multichannel DYPSSA smoothing function	120
4.12	Photograph of microphone array	121
4.13	Identified GCIs in reverberation	122
4.14	Detection rate vs. reverberation time for GCI detection algorithms	123
4.15	Identification accuracy vs. reverberation time for GCI detection algorithms	124
5.1	DDVSM model training	129
5.2	DDVSM Model Analysis-Synthesis	130
5.3	Fisher's Discriminant as a function of M classes.	132
5.4	Five MFCC-derived prototype voice source waveforms	134
5.5	The cumulative energy of the subset of $K' < K$ principal components.	136
5.6	Principal components for voice source data matrix	137
5.7	The mean vector and the first four principal components.	138
5.8	PCA/GMM Prototypes	139
5.9	Mean glottal excitation and least-square preemphasis filter	142
5.10	Estimated glottal excitation waveforms with different types of preemphasis	143
5.11	Energy decay curve for $p_{enh}(n)$	143
5.12	PCA Resynthesis	144
5.13	Feature-domain speech signal analysis	145
5.14	Probability of a speaker given mixture a component	146
6.1	Microphone Array	150
6.2	System diagram of multichannel dereverberation algorithm	152
6.3	Segmental SNR for dereverberation algorithms	155
6.4	BSD for dereverberation algorithms	155
6.5	The PSOLA Algorithm	158
6.6	Errors with treating unvoiced sounds as voiced with PSOLA time expansion	159
6.7	Errors with treating unvoiced sounds as voiced with PSOLA time compression	160
6.8	State probabilities for voiced, unvoiced and silent speech	161
6.9	MOS scores as a function of time scale modification	163
6.10	MOS confidence intervals as a function of time scale modification	163

6.11	Speech spectra for unvoiced and voiced speech	165
6.12	Spectral mirroring for unvoiced speech	166
6.13	Spectral mirroring for voiced speech	167
6.14	Artificial bandwidth extension source estimation	169
6.15	System diagram for artificial bandwidth extension synthesis	170
6.16	MOS results for artificial bandwidth extension algorithms	172
A.1	SIGMA with varying group delay length for male speech	203
A.2	SIGMA with varying group delay length for female speech	204
B.1	Sixteen MFCC/GMM-derived voice source prototypes	206
B.2	Sixteen PCA/GMM-derived voice source prototypes	207

List of Tables

3.1	GCI detection with EGG signals from APLAWD	90
3.2	GCI detection with EGG signals from APLAWD	90
3.3	GCI detection with EGG signals from SAM	91
3.4	GOI detection with EGG signals from SAM	91
4.1	GCI/GOI detection with speech signals from APLAWD	116
4.2	GCI detection with reverberant speech signals from APLAWD	124
6.1	Mean MOS scores for timescale-modified speech	163
6.2	Wideband envelope estimator configurations	171

Statement of Originality

As far as the author is aware, the following aspects of the thesis are believed to be original contributions:

1. Investigation into alternative methods for the detection of glottal closure and opening instants in voiced speech from Electroglottograph signals, resulting in an algorithm that employs feature extraction and Gaussian Mixture Modelling to obtain a measured accuracy of 99.59% against a hand-labelled database.
2. Investigation into alternative methods for the detection of glottal closure and opening instants in voiced speech from speech signals, resulting in an algorithm that employs voice source estimation, N -best Dynamic Programming and a novel voicing detector to obtain a measured accuracy of 99.84% against a hand-labelled database.
3. Extension of the existing DYPSA algorithm to a multichannel variant for the detection of glottal closure instants from reverberant voiced speech, achieving a measured accuracy of up to 18% greater than DYPSA in a reverberation time of 0.5 s.
4. Conceiving and implementing a framework for data-driven modelling of the voice source signal by segmenting the voice source into individual cycles using glottal closure instants from techniques developed in this work, then analysing with machine learning techniques.
5. Applying data-driven voice source modelling techniques to derive an improved LPC preemphasis filter, a coding scheme, speaker identification, voice source classification and artificial bandwidth extension of telephone signals.
6. Practical implementation of the SMERSH dereverberation algorithm, using glottal

closure instants derived with the multichannel technique presented in this work, to apply temporal and spatial averaging of the speech signal that suppresses unwanted reverberation components and improves SNR by up to 5 dB.

7. Development of a practical speech time-scaling algorithm, combining the Pitch-Synchronous Overlap-Add (PSOLA) algorithm and voiced/unvoiced silence detection to significantly improve subjective scores when compared with standard PSOLA.

Abbreviations

ABWE	Artificial Bandwidth Extension
ACR	Absolute Category Rating
AMR	Adaptive-Multi Rate
AR	Autoregressive
ARX	Autoregressive Exogeneous Input
BSD	Bark Spectral Distortion
BWE	Bandwidth Extension
CELP	Code-Excited Linear Prediction
CODEC	Coder/Decoder
CQ	Closed Quotient
DCR	Degradation Category Rating
DDVSM	Data-Driven Voice Source Modelling
DFT	Discrete Fourier Transform
DP	Dynamic Programming
DSB	Delay-and-Sum Beamformer
DSP	Digital Signal Processing
DTFT	Discrete-Time Fourier Transform
DWT	Discrete Wavelet Transform
DYPSA	DYnamic programming Phase Slope Algorithm
EGG	Electroglottogram
EM	Expectation Maximization
FA	False Alarm
FAT	False Alarm Total

FIR	Finite Impulse Response
FT	Fourier Transform
GCC-PHAT	Generalized Cross-Correlation PHASE Transform
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Modelling
GOI	Glottal Opening Instant
HQTx	High Quality Time of Excitation
IAIF	Iterative Adaptive Inverse Filtering
IDFT	Inverse Discrete Fourier Transform
IFT	Inverse Fourier Transform
IIR	Infinite Impulse Response
ITU	International Telecommunication Union
LCQA	Low-Complexity Quality Assessment
LP	Linear Prediction
LPC	Linear Predictive Coding
Lx	Laryngeal excitation
MA	Moving Average
MC-DYPSA	Multichannel DYPSA
MFCCs	Mel-Frequency Cepstrum Coefficients
MLS	Maximum Length Sequence
MOS	Mean Opinion Score
OQ	Open Quotient
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PSOLA	Pitch-Synchronous Overlap-Add
RAPT	Robust Algorithm for Pitch Tracking
RIR	Room Impulse Response
RTF	Room Transfer Function
SIGMA	Singularity in EGG by Multiscale Analysis

SMERSH	Spatiotemporal Method for Enhancement of Reverberant Speech
SNR	Signal-to-Noise Ratio
SSNR	Segmental Signal-to-Noise Ratio
SWT	Stationary Wavelet Transform
TDoA	Time Delay of Arrival
Tx	Time of excitation
TxGen	Time of Excitation Generator
UT	Unvoiced/Transition
VUS	Voiced/Unvoiced/Silence
YAGA	Yet Another GCI Algorithm

Mathematical Notation

General Notations

a	Scalar Quantity
$\mathbf{a}$	Vector Quantity
$\mathbf{A}$	Matrix Quantity
$a(t)$	Function of a discrete variable t
a_t	Function of a finite discrete variable t
$A(\omega)$	Fourier transform of a discrete function $a(t)$
$A(z)$	z -transform of a discrete function $a(t)$

Operators

$a * b$	Linear convolution
$\mathbf{A}^T$	Non-conjugate matrix transpose
$\mathbf{A}^{-1}$	Matrix inverse
a^*	Complex conjugate
$\Re\{\cdot\}$	Real part of complex value
$\Im\{\cdot\}$	Imaginary part of complex value
$\ \cdot\ _p$	L^p -norm
$ \cdot $	Absolute value
$\lceil \cdot \rceil$	Ceiling
$\lfloor \cdot \rfloor$	Floor
$\text{round}(\cdot)$	Nearest integer
$\max(\cdot)$	Largest value
$\min(\cdot)$	Smallest value
$\arg \max(\cdot)$	Argument of the maximum
$\uparrow_{\alpha}^{\beta}$	Resampling factor β/α
$\mathcal{F}$	Discrete-Time Fourier Transform
$a(n) \iff A(z)$	Equivalence of transform representations
$E\{\cdot\}$	Expected value
$p(\cdot)$	Discrete probability
$p(x y)$	Conditional probability of event x given event y
$f(\cdot)$	Probability density function
$\{\cdot\}^+$	Positive-going half-wave rectification
$\{\cdot\}^-$	Negative-going half-wave rectification
$S^+\{\cdot\}$	Schmitt trigger
$W_{2^j}\{\cdot\}$	Stationary Wavelet Transform at scale j

Symbols and Variables

a_k	Linear Predictive Coding (LPC) coefficients
a_j^s	SWT approximation coefficients at scale j
$A(z)$	z -transform of LPC coefficients
A_k	Segmental vocal tract area
b	Bark scale
$b_{m,k}$	LPC coefficients for channel m
b_k	Best-fit LPC coefficients for multiple channel
$B(z)$	z transform of best-fit LPC coefficients
$B(z)$	Numerator of ARMA model (Chap. 2)
$\mathcal{B}(r)$	Bark spectrum for frame r
c	Speed of sound
$\mathbf{c}$	Autocorrelation vector
$\mathbf{c}_r$	Feature vector for cycle r
$c_A(r)$	Speech waveform similarity cost at GCI r
$c_P(r)$	Pitch deviation cost at GCI r
$c_J(r)$	Projected candidate cost at GCI r
$c_F(r)$	Normalized energy cost at GCI r
$c_S(r)$	Ideal phase-slope deviation cost at GCI r
$c_I(r)$	Interchannel correlation cost at GCI r
$c(n)$	Complex cepstrum
$C(\omega)$	DTFT of complex cepstrum
$\tilde{c}(n)$	Liftered complex cepstrum
$\mathbf{c}_\Omega$	Cost vector for GCI set Ω at candidate r
$\mathbf{c}_\Omega(r)$	Cost vector from at GCI r
$\mathcal{C}_r$	Closed phase r
$d(n)$	Interchannel GCI correlation function at sample n (Chap. 4)
$d(n)$	Time-domain impulse train
$D(z)$	Frequency-domain impulse train
d_j^s	SWT detail coefficients at scale j
$e(n)$	Time-domain LPC residual
$E(z)$	z -domain LPC residual
$\bar{\mathbf{e}}_r$	LPC residual vector for r th glottal cycle at the output of a DSB
$\hat{\mathbf{e}}_r$	Enhanced LPC residual vector for r th larynx cycle at the output of a DSB
$\bar{e}(n)$	Delay-and-Sum Beamformer (DSB) prediction residual at sample n
$e_m(n)$	Reverberant LPC residual at channel m and sample n
$\hat{e}(n)$	Enhanced LPC residual
$e^{nb}(n)$	Narrowband LPC residual
$e^{wb}(n)$	Wideband LPC residual
E_e	LF parameter: Amplitude at time T_e
f_0	Fundamental glottal frequency
f_{max}	Maximum f_0
f_{min}	Minimum f_0
f_s	Sampling frequency
$\{F\}$	Speech frame
$F(n)$	Frobenius norm at sample n

F_m	Fisher ratio for m classes
f_X	PDF for variable X
F	Total number of features
$\mathbf{f}_r$	Feature vector for cycle r
$\mathbf{F}$	Feature matrix
$\mathbf{F}^{hb}$	Highband parameter matrix
$\mathbf{f}_r^{nb}$	Narrowband feature vector
$\mathbf{f}_r^{wb}$	Wideband feature vector
$g_m(n)$	Impulse train of Glottal Closure Instant (GCI)s for channel m
$\bar{\mathbf{g}}_r$	Equalization filter for glottal cycle r
$\bar{g}_{r,n}$	Equalization filter for glottal cycle r at sample n
$\hat{\mathbf{g}}_r$	Least-squares estimate of equalization filter vector for glottal cycle r
$\bar{g}(n)$	Impulse train of GCIs
$g(n)$	Time-domain glottal pulse volume velocity
$G(z)$	z -domain glottal pulse volume velocity
$g_D(n)$	Time-domain glottal pulse voice source waveform
$G_D(z)$	z -domain glottal pulse voice source waveform
$G_{n\hat{\epsilon}}(z)$	z -transform of equalization filter for glottal cycle r
$g_j(n)$	SWT detail filter at scale j
$h_d(n)$	Direct-path impulse response
$h_m(n)$	Room Impulse Response (RIR) at channel m and sample n
$\mathbf{h}$	Vector of 1s
$\mathbf{h}_m$	RIR vector at channel m
$h_j(n)$	SWT approximation filter at scale j
I	Number of neighbouring samples
$\mathcal{I}$	Total number of frames
$\mathbf{I}$	Identity matrix
$I(n)$	Ideal group delay slope
j	SWT scale
J	Total SWT scales
k	Filter tap / discrete segment of vocal tract
k	DTFT bin (Chapter 3)
K_G	Gain coefficient of two-pole glottal model
K_R	Gain coefficient of lip radiation filter
K_V	Gain coefficient of all-pole vocal tract model
K	Frobenius norm window length (samples) (Chap. 4)
k	Basis vector index (Chap. 5)
K	Speech cycle normalization length (samples) / total transform spectra (Chap. 5)
K'	Truncated voice source transform spectra
l	Index of finite speech segment
L	Length of finite speech segment (Chap. 5, 6)
$l(n)$	Liftering window
L	RIR length (Chap. 2)
L	Vocal tract length (Chap. 2)
L_g	Equalization filter length
M	Total number of microphones
m	Class index (Chap. 5)
m	Microphone index

m	GMM mean vector
$\mathcal{M}$	Voice source model
M	Total number of classes (Chap. 5)
$N_{\{1,2\}}$	Rosenberg model parameters
n	Sample index
n_r^c	Sample index of GCI r
$\check{n}_r^c$	Sample index of candidate GCI r
n_r^{cref}	Sample index of reference GCI r
$n_{r,m}^c$	Sample index of GCI r on channel m
n_r^o	Sample index of GOI r
$\check{n}_r^o$	Sample index of candidate GOI r
n_r^{oref}	Sample index of reference GOI r
$n_{r,m}^o$	Sample index of GOI r on channel m
n_0	Lifting window boundary
n'	Finite set of speech cycles
$\tilde{n}_i^c$	Timescale modified GCIs
N	Speech segment length (samples)
N_{ov}	Crossfade length (samples)
$O(n)$	Order of complexity n
p	Prediction order
P	Probability matrix
$p(n)$	Time-domain preemphasis filter
$P(z)$	z -domain preemphasis filter
$p_{enh}(n)$	Time-domain enhanced preemphasis filter
$\mathbf{p}_{enh}$	Vector time-domain enhanced preemphasis filter
$P_{enh}(z)$	z -domain enhanced preemphasis filter
$p^{\{+,-\}}$	Rectified multiscale product
P_m	Pneumotachograph acoustic pressure
$P(z)$	z -domain preemphasis filter
$P\{\omega\}$	Prior probability for class ω
$q_n(r)$	Windowed stationary wavelet transform beginning at sample n
$Q_n(k)$	DTFT of $q_n(r)$
$\tilde{Q}_n(k)$	DTFT of $rq_n(r)$
Q_r^c	Closed quotient for candidate r
Q_r^o	Open quotient for candidate r
Q_{max}^o	Maximum open quotient
Q_{min}^o	Minimum open quotient
Q_E	Sum-square AR model error
r	GCI/GOI index
$\hat{r}$	Selected GCI/Glottal Opening Instant (GOI) index
r	Reflection coefficient (Chap. 2)
R	Number of voice source cycles
R_k	Autocorrelation at lag k
R^c	Number of detected GCIs
$\check{R}^c$	Number of GCI candidates
R^{cref}	Number of reference GCIs
R^o	Number of detected GOIs

$\check{R}^o$	Number of GOI candidates
R^{ref}	Number of reference GOIs
R_m^c	Number of GCIs at channel m
$\check{R}_m^c$	Number of GCI candidates at channel m
R_m^o	Number of GOIs at channel m
$\check{R}_m^o$	Number of GOI candidates at channel m
R_s	Number of voice source cycles from speaker s
$R(z)$	z -domain lip radiation characteristic
$\hat{R}(z)$	z -domain estimate of lip radiation characteristic
$R_{x,y}(\tau)$	Cross-correlation between variables x and y at lag τ
$\mathcal{R}$	Subset of Glottal Closure Instants (GCIs)
$\mathbf{R}_{xx}$	Autocorrelation matrix from x
$\mathbf{r}_{xy}$	Cross-correlation vector formed from x and y
$s(n)$	Input speech signal at sample n
$S(z)$	z -domain input speech signal
$S(\omega)$	DTFT of input speech signal
$s'(n)$	Time-aligned speech signal at sample n
$\mathbf{s}(n)$	Input speech signal vector
$\hat{s}(n)$	Enhanced / estimated speech signal
$\hat{\mathbf{s}}(n)$	Vector of enhanced / estimated speech
$\tilde{s}$	Windowed frame of speech
$s^{nb}(n)$	Narrowband speech signal
$S^{nb}(z)$	z -domain narrowband speech signal
$s^{wb}(n)$	Wideband speech signal
$S^{wb}(z)$	z -domain wideband speech signal
$\hat{s}^{lb}(n)$	Estimated lowband speech signal
$\hat{s}^{hb}(n)$	Estimated highband speech signal
$\hat{s}^{wb}(n)$	Estimated wideband speech signal
$\mathcal{S}$	Silent class
t	Continuous time index (s)
t_{max}	Maximum glottal period (s)
t_{min}	Minimum glottal period (s)
T	Sampling period
T_{60}	Reverberation time (s)
$\mathbf{T}^{lb}$	Lowband time envelope parameter matrix
$\mathbf{T}^{hb}$	Highband time envelope parameter matrix
$u_G(n)$	Time-domain glottal volume velocity
$U_G(z)$	z -domain glottal volume velocity
$\hat{u}_G(n)$	Estimate of time-domain glottal volume velocity
$\hat{U}_G(z)$	z -domain estimate of glottal volume velocity
$u_D(n)$	Time-domain glottal pressure (voice source)
$U_D(z)$	z -domain voice source
$U_D(\omega)$	DTFT of voice source
$\hat{u}_D(n)$	Estimate of time-domain glottal pressure (voice source)
$\hat{U}_D(z)$	z -domain estimate of voice source
$u_L(n)$	Time-domain lip volume velocity
$U_L(z)$	z -domain lip volume velocity

$\hat{u}_L(n)$	Estimate of time-domain lip volume velocity
$\hat{U}_L(z)$	z -domain estimate of lip volume velocity
$u_D^{wb}(n)$	Time-domain wideband voice source
$U_D^{wb}(z)$	z -domain wideband voice source
$u_D^{nb}(n)$	Time-domain narrowband voice source
$U_D^{nb}(z)$	z -domain narrowband voice source
u_k	Forward volume velocity at vocal tract segment k
$u_D^{Rosenberg}(n)$	Rosenberg model of glottal volume velocity
$u_D^{LF}(n)$	Liljencrants-Fant model of glottal volume velocity
$\mathbf{u}_r$	Normalized voice source cycle pair centred at GCI r
$\hat{\mathbf{u}}_r$	Estimated normalized voice source cycle pair centred at GCI r
$\tilde{\mathbf{u}}_r$	Preconditioned, normalized voice source cycle pair centred at GCI r
$\bar{\mathbf{u}}$	Mean normalized white source signal
$\bar{\mathbf{u}}_m$	Prototype voice source cycle vector for class m
$\mathbf{u}_r^{wb}$	Wideband normalized voice source cycle pair centred at GCI r
$\mathbf{u}_r^{nb}$	Narrowband normalized voice source cycle pair centred at GCI r
$\bar{\mathbf{u}}_{nb,m}^{wb}$	Prototype voice source vector for class m , trained on NB params & WB speech
$\bar{u}_i$	Left-singular vectors for singular value σ_i
$\mathcal{U}$	Unvoiced class
U_m	Pneumotachograph volume velocity
$U(z)$	Generic volume velocity
$\mathbf{U}$	Left-singular matrix
$\mathbf{U}$	Normalized voice source cycle pair matrix
$\hat{\mathbf{U}}$	Estimated normalized voice source cycle pair matrix
$\bar{\mathbf{U}}$	Prototype voice source cycle matrix
$\mathbf{U}^{wb}$	Wideband normalized voice source cycle pair matrix
$\mathbf{U}^{nb}$	Narrowband normalized voice source cycle pair matrix
$\bar{\mathbf{U}}_{nb}^{wb}$	Prototype voice source matrix, trained on NB params & WB speech
v_g	Backward volume velocity at the glottis
v_k	Backward volume velocity at vocal tract segment k
$\bar{v}_i$	Right-singular vectors for singular value σ_i
$\mathcal{V}$	Voiced class
$V(z)$	z -domain vocal tract transfer function
$\bar{V}(z)$	z -domain vocal tract transfer function with gain and propagation terms
$V(\omega)$	DTFT of vocal tract transfer function
$\hat{V}(z)$	z -domain estimate of vocal tract transfer function
$\mathbf{V}$	Right-singular matrix (Chap. 2)
$\mathbf{V}$	PCA component matrix (Chap. 5)
$\mathbf{V}'$	Truncated PCA component matrix
w_m	GMM weight for class m
$w(n)$	Windowing function
$w_r(n)$	Aggregate crossfade windowing function
$w^{in}(n)$	Fade-in window
$w_r^{in}(n)$	Fade-in window for cycle r
$w^{out}(n)$	Fade-out window
$w_r^{out}(n)$	Fade-out window for cycle r
$\mathbf{W}$	Diagonal weighting matrix

$x_m(n)$	Observed speech signal at microphone m and sample n
$X_m(\omega)$	DTFT of observed speech signal at channel m
$x_m(t)$	Continuous-time observed speech signal at microphone m and sample n
$\mathbf{x}_m(n)$	Observed speech vector at microphone m and sample n
$\bar{x}(n)$	Output of DSB at sample n
$x_n(l)$	Finite segment of speech beginning at sample n
$X_n(k)$	Discrete-Time Fourier Transform (DTFT) of $x_n(l)$
$\tilde{X}_n(k)$	DTFT of $rx_n(l)$
$x_{ref}(n)$	Reference channel
$X_{ref}(\omega)$	Fourier Transform (FT) of reference channel
$\mathbf{x}_i$	Feature vector at frame index i
$\mathbf{x}_r$	Zero-mean voice source cycle r
$\hat{\mathbf{x}}_r$	Estimated zero-mean voice source cycle r
$\mathbf{X}$	Zero-mean voice source cycles
$\hat{\mathbf{X}}$	Estimated zero-mean voice source cycles
x_f^{nb}	Narrowband feature in frame f
$\mathbf{Y}$	General orthogonal voice source transform matrix
$\mathbf{Y}'$	Truncated general orthogonal voice source transform matrix
Z_m	Pneumotachograph acoustic impedance
$Z(x)$	Vocal tract impedance as a function of length x
$\mathbf{z}_r$	PCA spectra for cycle r
$\mathbf{z}'_r$	Truncated PCA spectra for cycle r
$\mathbf{Z}$	Matrix of PCA spectra
$\mathbf{Z}'$	Truncated matrix of PCA spectra
α	LF parameter: determines ratio of E_e to peak amplitude
$\boldsymbol{\alpha}$	Vector of AR coefficients
β	LF parameter: Exponential time constant for return phase
$\gamma(n)$	Group delay function at sample n
$\bar{\gamma}(n)$	Mean group delay value at sample n
Γ_r	Class set producing highest likelihood for cycle r
δ	GCI/GOI error (samples)
$\delta(n)$	Unit impulse function at sample n
Δ_P	Pitch deviation
ϵ	Incremental value
ζ	Pitch deviation cost mapping coefficient
ζ_r^{in}	Fade-in offset for cycle r
ζ_r^{out}	Fade-out offset for cycle r
$\eta(n)$	Electroglottograph signal
$\Theta(\omega)$	DTFT phase at frequency ω
ι	Ideal phase slope function deviation window width (samples)
κ_r	Amplitude normalization coefficient for cycle r
κ_m	Amplitude normalization coefficient for class m
$\kappa_{m,r}$	Amplitude normalization coefficient for class m , cycle r
$\boldsymbol{\lambda}$	Vector of DYPSA weighting factors
λ_k	Eigenvalue k
μ	Lip radiation zero (Chap. 2)
μ	GCI/GOI bias (samples)
μ_m	GMM cluster mean m

μ_m^{nb}	GMM narrowband cluster mean m
ν	Waveform similarity threshold
ξ	GOI open quotient tolerance
ξ_s	Event of speaker s talking
ρ	Density of air (Chap. 2)
ρ	GOI dynamic programming state variable (Chap. 4)
σ	GCI/GOI identification accuracy (samples)
σ_i	Singular value i
Σ_m	GMM covariance matrix for class m
Σ_m^{nb}	GMM narrowband covariance matrix for class m
τ_m	Time Delay of Arrival (TDoA)
$\tau_n(k)$	Group delay at time n samples
$\hat{\tau}_{GCC}$	GCC TDA estimate
$\hat{\tau}_{GCC}$	Estimated TDoA by GCC-PHAT
$\Upsilon(n)$	GCI smoothing function at sample n
$\phi_{i,j}$	Covariance at lag i and j
Φ	Covariance matrix
φ_r	GMM-based voice source decomposition weight vector for cycle r
$\varphi_{m,r}$	GMM-based voiced source decomposition for class m , cycle r
χ	Pitch deviation coefficient
ψ_k	Basis vector k
ψ'_k	Truncated basis vector k
Ψ	Basis matrix
Ψ'	Truncated basis matrix
ω_0	LF parameter: curvature preceding glottal closure
ω	Continuous frequency variable
ω_k	Weighting coefficient at tap k
ω	VUS class
ω_m	Voice source class m
ω_m^{nb}	Narrowband voice source class m

Chapter 1

Introduction

1.1 Context of Work

SPEECH processing is a field of engineering which is integral to modern living. Our understanding of the nature of the human voice and the models used to describe it have made commonplace such applications as mobile telephony, speech recognition, speech enhancement and speech synthesis. Central to speech processing is the linear model of speech [1] that describes voiced speech as a linear combination of three components: the voice source, the vocal tract and lip radiation. The voice source is a pseudoperiodic signal produced by the the glottis as it rapidly opens and closes. This signal is spectrally filtered by the vocal tract and lip radiation characteristic to produce sounds which are interpreted by a listener as voiced phonemes.

1.1.1 Voice Modelling

The characteristics of the vocal tract and lip radiation are well-understood, for which there are many accurate and compact representations. However, the voice source signal is less well-understood and few models exist that can reproduce the whole gamut of waveforms with a compact set of parameters. As a pseudoperiodic process, the key to devising good models is the detection of period boundaries. The Glottal Closure Instants (GCIs) are primary impulsive instants of excitation which can be detected, forming the basis

of Glottal-Synchronous Speech Processing. The Glottal Opening Instants (GOIs) are secondary excitation instants which are useful in closed-phase analysis [2] and analysing pathological speech [3]. Detection of GCIs and GOIs from both Electroglottogram (EGG) measurements and speech recordings are challenging problems that are investigated in detail in this thesis.

Existing models of the voice source are motivated by either parametric curve-fitting of observations made by speech scientists, physical descriptions of the glottis or stochastic codebooks. Although these models have been successfully applied to real-world problems, the methods presented in this thesis for accurate detection of GCIs enables deeper investigation into voice source characteristics that can be found in large training databases of speech. The concept of Data-Driven Voice Source Modelling (DDVSM) is presented, setting out a framework for which machine learning techniques may be applied to determine the degrees of freedom of the voice source. The compact and accurate models produced by DDVSM are shown to have potential in a number of real-world applications.

1.1.2 Application Context

Provided GCIs can be accurately determined, the following are examples of real-world speech processing applications investigated in this thesis:

Dereverberation and noise reduction Additive noise sources and sound reflected from walls, termed reverberation, can distort a speech signal by reducing intelligibility and perceived quality [4]. Section 6.2 describes a technique that performs spatial averaging on multichannel observations and temporal averaging on neighbouring glottal cycles. In doing so, signal components from voiced speech are enhanced and reverberation/additive noise components are attenuated, improving the perceived quality of the recorded speech. A GCI detector that is robust to reverberation is also required and is discussed in Section 4.5.

Time-Scale Modification The slowing or speeding up of speech using dissimilar record and playback speeds is disadvantageous because both duration, pitch and formant

structure are affected, resulting in very unnatural-sounding speech. Copying or removing fixed frames of speech generates noticeable artefacts as the pseudoperiodicity of the processed speech is not preserved. In Section 6.3, natural-sounding time-scale modification is achieved by instead copying or removing individual cycles of speech with particular reference to the importance of voicing detection.

Artificial Bandwidth Extension Speech transmission channels limit the audio bandwidth of the transmitted speech to reduce data bandwidth; most telephony standards limit audio to $\sim 300 \text{ Hz} - 3.4 \text{ kHz}$. Wideband speech is considered to be $\sim 50 \text{ Hz} - 7 \text{ kHz}$. Many methods exist for estimating the upper extension band, where spectral envelope is of high importance, whereas relatively few exist for the low extension band for which temporal characteristics can be shown to be more significant [5]. An implementation of DDVSM with multiple audio bandwidths is applied in Section 6.4 to accurately estimate the low extension band.

Coding and Speaker Identification Further applications of DDVSM include coding and speaker identification, discussed in Section 5.5. The ability of data-driven models to capture features of the voice source in a compact set of parameters is the basis for an efficient speech coding scheme. The identification of clustering within model parameters can be shown to sometimes be speaker-specific and may aid existing speaker identification algorithms.

1.2 Research Statement and Thesis Structure

1.2.1 Research Statement

The aim of this thesis is to develop methodologies for the detection of GCIs and GOIs from speech and EGG signals, to exploit them in existing real-world problems and to define new models of the glottal waveform with data-driven techniques.

1.2.2 Thesis Structure

The content of this thesis is as follows. In Chapter 2, the fundamentals of glottal-synchronous speech processing are reviewed. This includes a description of the physical processes involved in producing speech signals, the models available for describing their behaviour, and an overview of the advantages and challenges posed by glottal synchronous speech processing.

Chapter 3 discusses the invasive EGG technique for the measurement of glottal electrical impedance. The SIGMA algorithm for the detection of reference GCIs and GOIs from EGG signals and is evaluated in detail.

Chapter 4 discusses existing techniques for the detection of GCIs and GOIs from speech signals. The DYPSA algorithm is reviewed in detail and the YAGA and Multi-channel DYPSA algorithms are presented and evaluated for the detection of GCIs and GOIs in clean and reverberant recordings respectively.

In Chapter 5, Data-Driven Voice Source Modelling is proposed as a novel application of glottal synchronous speech processing. Building upon the foundations of the preceding chapters, voiced speech excitation is estimated from a large speech database. The excitation is then segmented and analysed from which data-driven models are derived. The application of the technique to analysis, coding and speaker identification is demonstrated.

In Chapter 6, glottal-synchronous techniques are applied to existing applications in speech processing, including dereverberation, time-scale modification and Artificial Bandwidth Extension. Various evaluation techniques are used to demonstrate the advantages of the methodology employed. The application to real-world problems highlights some of the challenges faced and the measures taken to overcome them.

The thesis is concluded and further work is discussed in Chapter 7.

Chapter 2

Review

2.1 Introduction

THE theory of speech processing is discussed in this chapter. First the physiology of speech is reviewed. Sections 2.3–2.7 then investigate the linear source-filter model of speech, reviewing in detail models of the voice source and vocal tract filter and the estimation and measurement of glottal waveforms.

The second half of this chapter begins in Section 2.8 by reviewing the problem of detecting Glottal Closure Instants (GCIs) and Glottal Opening Instants (GOIs) from noninvasive and invasive measurements. Their use in an application context is discussed in Sections 2.9–2.11. A chapter summary is given in Section 2.12.

2.2 Physiology of Voice Production

All voiced sounds are produced by a *source*, or *excitation*, signal that is spectrally filtered by the vocal tract. This excitation is produced inside the larynx by the vocal folds, which consists of opposing ligaments that form a constriction at the top of the trachea as it joins the lower vocal tract, depicted in Figures 2.1 (a) and 2.2 (a). When air is expelled from the lungs at sufficient velocity through this orifice – usually referred to as the glottis – the air pressure causes the vocal folds to experience a separating force as depicted in Figure 2.2

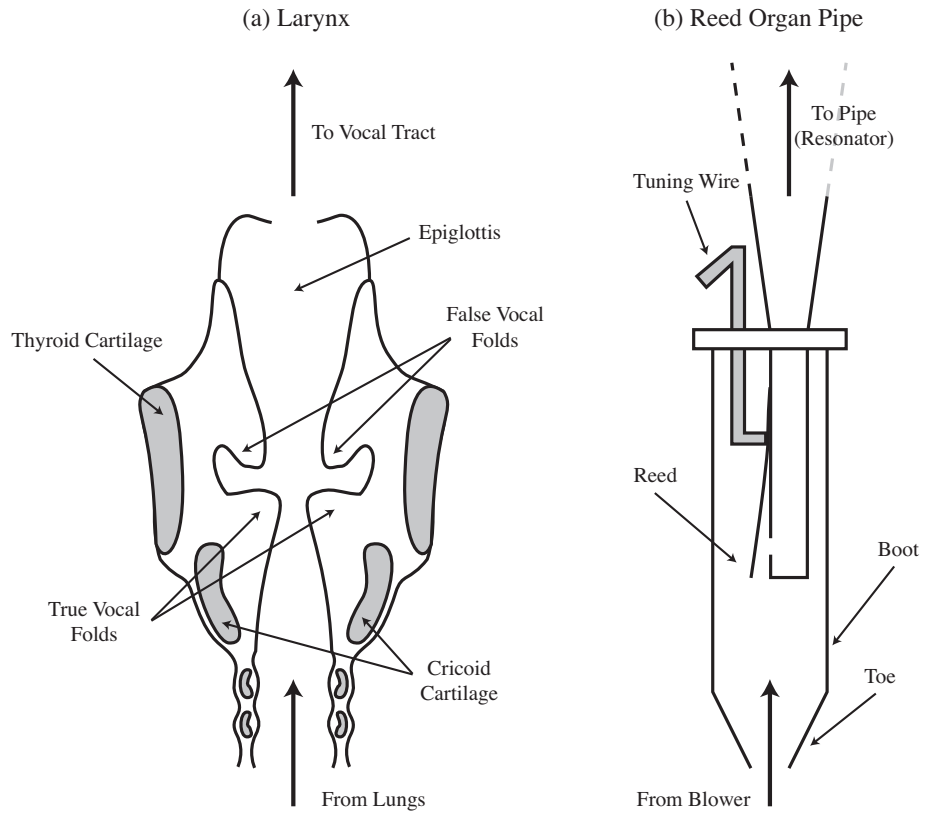


Figure 2.1: Cross section schematics of (a) larynx (after [6]) and (b) reed organ pipe. In (a), airflow from the lungs causes the vocal folds to vibrate, whose frequency is controlled largely by muscle tension. In (b), airflow from a blower causes a reed to vibrate, whose frequency is controlled by a tuning wire.

(b), (c). This instant of time is termed the Glottal Opening Instant (GOI). The vocal folds continue to open until equilibrium is reached between the separating force and the tension in the vocal folds, at which point the potential energy stored in the vocal folds causes them to begin to close in Figure 2.2 (d). When the vocal folds become sufficiently close, the suction effect of the Bernoulli force results in an abrupt closure at the Glottal Closure Instant (GCI) in Figure 2.2 (f). Elastic restoring forces during closure cause the cycle to repeat, producing a series of pulses termed *modal* voiced speech, the most common vocal register in conversational speech [7]. The ratio of open time with respect to glottal period is termed the Open Quotient (OQ) [8].

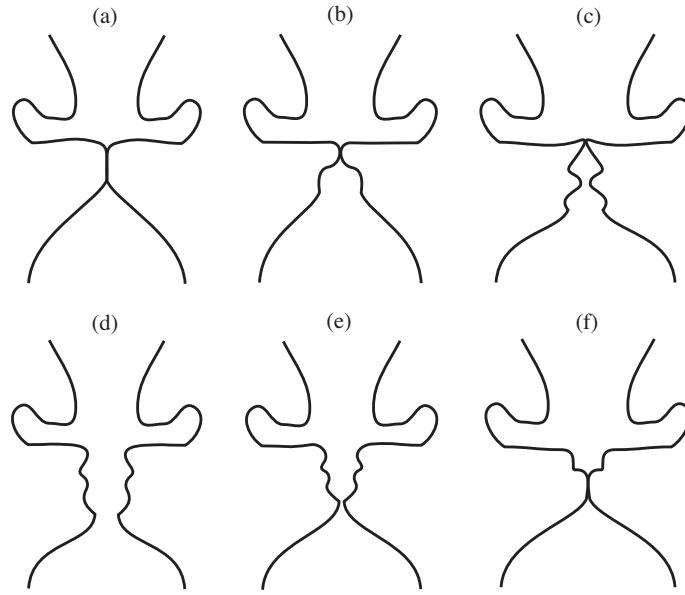


Figure 2.2: A sequence of cross-sections of the larynx for a complete glottal cycle (after [9]).

2.2.1 Physiology of the Vocal Tract

The vocal tract consists of the pharyngeal cavity, oral cavity and nasal cavity as depicted in Figure 2.3, which act as resonant tubes that spectrally filter the excitation. In addition to the vocal folds, the velum, tongue, teeth and lips form the *articulators* which are responsible for fine control of the sounds produced.

A number of wind instruments, including the oboe, clarinet and certain types of organ pipe, produce sounds with a similar process where an excitation signal is created with a reed that vibrates and excites a resonant tube. A reed organ pipe is shown in Figure 2.1(b). The vocal folds are excited by air from the lungs where the frequency of vibration, f_0 , is controlled by their size and tension. In the organ pipe, air is provided by a blower and frequency is fixed by the tension placed on the reed with a tuning wire. The vocal tract and resonant pipe both perform spectral shaping on the excitation signal as discussed in Section 2.3.1.

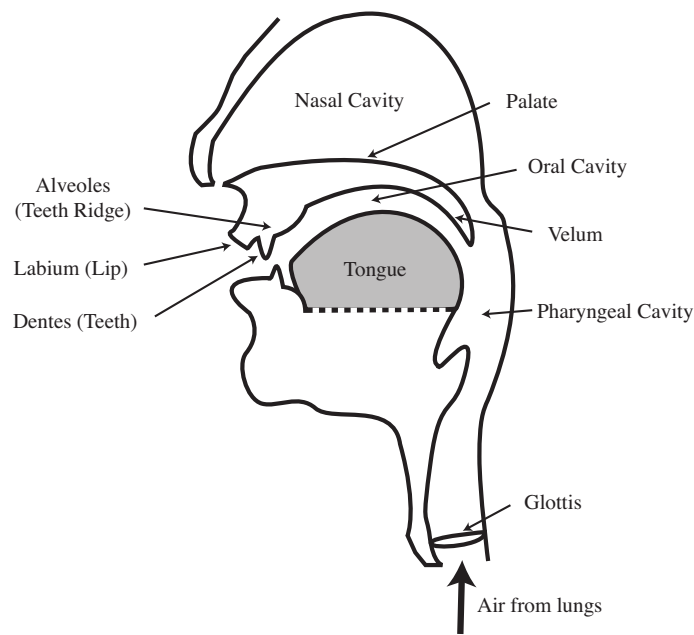


Figure 2.3: Schematic diagram of vocal tract physiology.

2.3 Models of Speech Production

Modelling of speech relies upon the measurement or estimation of the location of the articulators and their effect upon the sounds produced. Early measurements used X-ray video images of the head [10] which, combined with more recent linear prediction techniques reviewed in Section 2.6.1, have led to the determination of vocal tract shape for vowels [11]. An analysis-by-synthesis approach was further applied in the estimation of the vocal tract for fricatives [11]. Direct measurement of the articulators' positions is useful in certain areas of speech science but does not necessarily provide useful information about acoustic behaviour. Much interest has been shown in models that are motivated by the vocal tract's acoustic transfer function, notably the linear source-filter model [1] described in Section 2.3.1. Various nonlinearities have been suggested including the nonlinear interaction between the glottis and the vocal tract [12], and thermal loss/viscosity [13]. Other nonlinear models estimate parameters that do not correspond to physical attributes including nonlinear prediction [14] and AM-FM models [15].

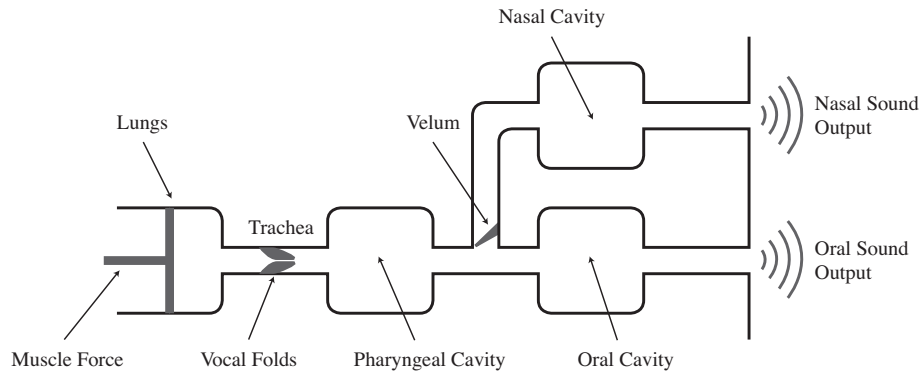


Figure 2.4: Block diagram of human speech production. The lungs force air through the vocal folds, causing them to vibrate. The glottal pulses are spectrally filtered by the vocal tract which consists of the pharyngeal, oral and nasal cavities. Sound is radiated from the nasal and oral cavities which can be approximated as a 1st-order differentiator.

2.3.1 The Linear Source-Filter Model

The linear source-filter model [1] is used throughout this thesis and is common in the wider context of speech processing. A simplified schematic figure of voiced speech production is shown in Figure 2.4 which shows the cavities as resonant chambers through which the glottal excitation signal passes. Combining the pharyngeal, oral and nasal cavities into a vocal tract filter, $V(z)$, voiced speech can be described with the linear source-filter model as three independent linear systems: the glottal volume velocity, $u_G(n) \iff U_G(z)$, the vocal tract filter, $V(z)$, and lip radiation filter, $R(z)$, producing a speech signal, $s(n) \iff S(z)$,

$$S(z) = U_G(z)V(z)R(z) \quad (2.1)$$

$$= U_L(z)R(z), \quad (2.2)$$

where $u_l(n) \iff U_L(z)$ is volume velocity at the lips. The glottal volume velocity, $U_G(z)$ [m^3s^{-1}], is analogous to electrical current. It can be factorized into an impulse train component, $d(n) \iff D(z)$, and a glottal pulse component, $g(n) \iff G(z)$, such that

$$S(z) = D(z)G(z)V(z)R(z). \quad (2.3)$$

Both the lip radiation filter and vocal tract can be modelled as acoustic transfer functions,

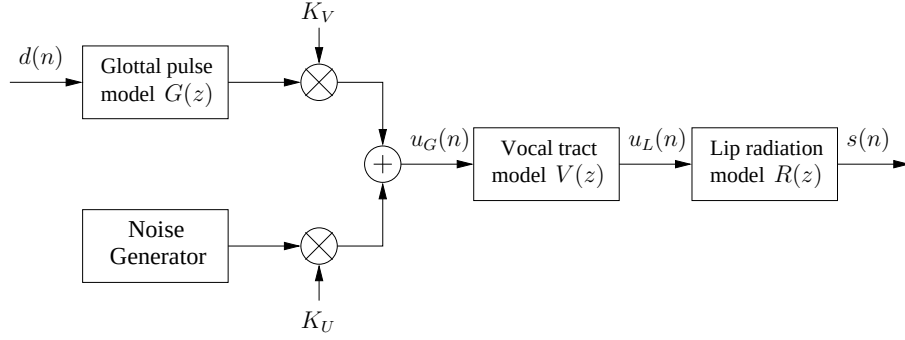


Figure 2.5: The source-filter model of speech. The vocal tract and lip radiation filter an excitation signal in the form of a glottal pulsetrain (voiced speech) and noise (unvoiced speech).

where the vocal tract is a lossless tube of varying cross-section [16] and lip radiation is an open-ended tube mounted in a spherical baffle (the head). Microphones used to measure speech signals are usually sensitive to acoustic pressure [Nm^{-2}], analogous to electrical potential, and not volume velocity. The lips can therefore be thought of as a device that converts the volume velocity behind the lips to a pressure waveform received by a microphone. An analysis of a piston with area A mounted in a spherical baffle is given in [17], which predicts a 1st-order highpass filter with a corner frequency

$$\frac{c}{\sqrt{4A}} \text{ Hz} \simeq 5 \text{ kHz}. \quad (2.4)$$

For sampling frequency $f_s < 20 \text{ kHz}$ a good approximation is [9]

$$R(z) = K_R(1 - z^{-1}), \quad (2.5)$$

with magnitude response

$$|R(z)|_{z=e^{j\omega}} = 2K_R \sin\left(\frac{\omega}{2}\right), \quad (2.6)$$

where K_R is a gain constant and lip-microphone propagation delay is ignored. The filter $V(z)$ is assumed invariant during the two-sample impulse response of $R(z)$, permitting the combination of the lip radiation and glottal pulse terms [18],

$$S(z) = D(z)G_D(z)V(z) = U_D(z)V(z), \quad (2.7)$$

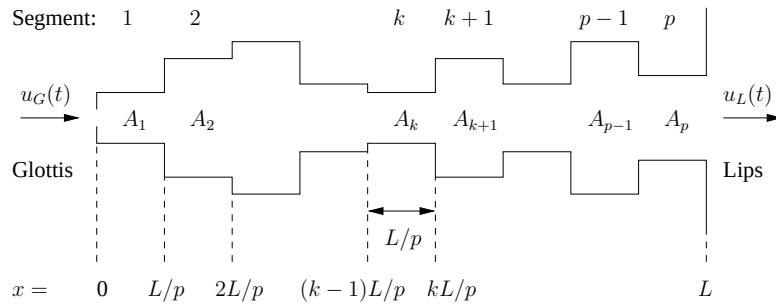


Figure 2.6: Discretized cross-section of the vocal tract. Changes in cross-section, and therefore acoustic impedance, cause the wavefront to be partially reflected. The reflections can be modelled as an all-pole filter, spectrally shaping the glottal excitation which is interpreted as different phonemes.

which both reduces the number of terms and converts the voice source volume velocity waveform, $U_G(z)$, to a pressure waveform, $u_D(n) \iff U_D(z)$, equivalent to a virtual microphone placed after the glottis. The signal $u_D(n)$ is usually referred to herein as the *voice source* although *glottal excitation*, *glottal pressure waveform* and *effective driving function* are common in the literature.

In the case of unvoiced speech, such as /s/, or mixed speech, such as /z/, noise-like excitation is produced not in the vocal folds but further towards the lips. The linear source-filter model of speech production is depicted in the time-domain in Figure 2.5 where K_V and K_U are gain terms for the voiced and unvoiced excitation respectively.

2.4 Modelling the Vocal Tract Transfer Function

The vocal tract can be modelled as a tube of varying cross-section, where it is assumed to be invariant for short periods of time. An all-pole model of the vocal tract is proposed in [16, 19] where its length L is divided into p equal-length segments of piecewise-constant cross-sectional area A_k as depicted in Figure 2.6.

The length of the vocal tract is ~ 17 cm for an adult male and ~ 14 cm for an adult female. The distance travelled by the acoustic wave in each section is L/p , equivalent to

half a sample period, $0.5T$. The total number of segments for an adult male is therefore

$$p = \frac{L}{0.5cT} \simeq \frac{f_s}{1000}, \quad (2.8)$$

where $c \simeq 340 \text{ ms}^{-1}$ is the speed of sound in air and f_s is sampling frequency.

The acoustic transfer function of the vocal tract can be found by analysing the superposition of volume velocity waves at the junction of neighbouring segments [18, 20]. Let u_k and v_k represent the forward and reverse volume velocities incident on segment k respectively. The total volume velocity in the segment is $u_k - v_k$ and total acoustic pressure is $(u_k + v_k) \times \rho c / A_k$, where ρ is the density of air and $\rho c / A_k$ is acoustic impedance. For a travelling wave, the volume velocity and pressure at the boundary of neighbouring segments is conserved,

$$\begin{aligned} u_k - v_k &= u_{k+1} - v_{k+1} \\ \frac{\rho c}{A_k}(u_k + v_k) &= \frac{\rho c}{A_{k+1}}(u_{k+1} + v_{k+1}). \end{aligned} \quad (2.9)$$

Solving for u_k and v_k , the incident and reflected waves at segment k can be found as a function of those at segment $k + 1$,

$$\begin{bmatrix} u_k \\ v_k \end{bmatrix} = \frac{1}{1 + r_k} \begin{bmatrix} z^{+\frac{1}{2}} & 0 \\ 0 & z^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 1 & -r_k \\ -r_k & 1 \end{bmatrix} \begin{bmatrix} u_{k+1} \\ v_{k+1} \end{bmatrix}, \quad (2.10)$$

$$= \frac{z^{\frac{1}{2}}}{1 + r_k} \begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} u_{k+1} \\ v_{k+1} \end{bmatrix}, \quad (2.11)$$

$$(2.12)$$

where $z^{\frac{1}{2}}$ is the propagation delay of a wave in each segment and the reflection coefficient r_k is given by

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}, \quad (2.13)$$

and is bounded on $-1 \leq r_k \leq 1$. Combining p cascaded segments yields

$$\begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \frac{z^{\frac{1}{2}p}}{\prod_{k=1}^p (1 + r_k)} \prod_{k=1}^p \begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} u_L \\ 0 \end{bmatrix}. \quad (2.14)$$

where it is assumed that $v_l = 0$. Considering the boundary with the glottis, it can be shown that [21]

$$u_G = \frac{2}{1 + r_G} [1, -r_G] \frac{z^{\frac{1}{2}p}}{\prod_{k=1}^p (1 + r_k)} \prod_{k=1}^p \begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} u_L \\ 0 \end{bmatrix}. \quad (2.15)$$

The resulting transfer function u_L/u_G is in the form

$$\bar{V}(z) = \frac{u_L}{u_G} = \frac{K_V z^{-\frac{1}{2}p}}{1 - \sum_{k=1}^p a_k z^{-k}}, \quad (2.16)$$

where K_V is a gain, $z^{-\frac{1}{2}p}$ is the acoustic propagation delay along the vocal tract and $a_k \iff A(z)$ form an all-pole (AR) filter. The gain and propagation terms are usually ignored for convenience such that

$$V(z) = \frac{\bar{V}(z)}{K_V z^{-\frac{1}{2}p}}. \quad (2.17)$$

The speech signal can then be calculated as a function of $u_D(n)$ and the vocal tract transfer function,

$$s(n) = u_D(n) + \sum_{k=1}^p a_k s(n - k). \quad (2.18)$$

which is a time-domain formulation of (2.7). Estimating the coefficients a_k from speech signals is important in speech analysis and is discussed in Section 2.6.1. Figure 2.7 shows the filter coefficients plotted in the z -domain and their frequency-domain transfer function for a frame of speech.

Assumptions in the All-Pole Vocal Tract Model

The all-pole model of the vocal tract is valid providing the following assumptions hold:

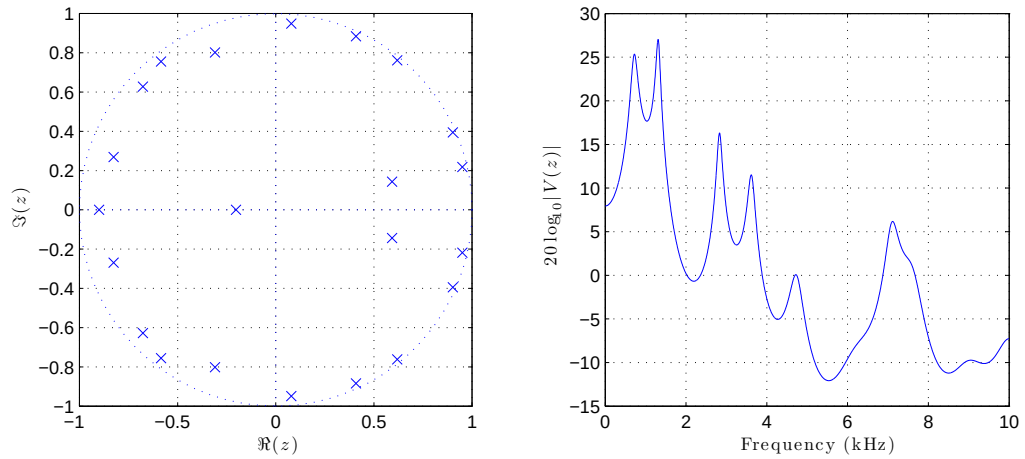


Figure 2.7: Pole plot and frequency response of the vocal tract for a frame of voiced speech.

1. Cross-sectional area is constant within a tube section.
2. The tube is lossless: the walls are rigid and there is no turbulent flow.
3. Longitudinal plane-wave propagation: sound pressure is constant throughout the cross-section.
4. Linear separability of the glottis and vocal tract.
5. Nasal cavity coupling.

The assumption of constant cross-sectional area within a tube section is valid providing the sampling rate is sufficiently high. The assumption of losslessness is particularly difficult to circumvent as the energy absorbed by the walls and viscous heating in the vocal tract cannot be easily measured. The longitudinal plane-wave assumption can be circumvented by the applying and solving the Navier-Stokes equations for a known vocal tract shape, although the simplicity of a linear all-pole model is replaced with a complicated nonlinear one. For frequencies < 3 kHz, where the majority of speech energy is contained, sound wavelengths are long compared with the tube diameter so the plane-wave assumption is valid [9]. The assumption of linear separability of glottis and vocal tract is dependent upon application; linear separability has played an important role in producing a working speech model, although physical models of the glottal waveforms reviewed in Section 2.5.2

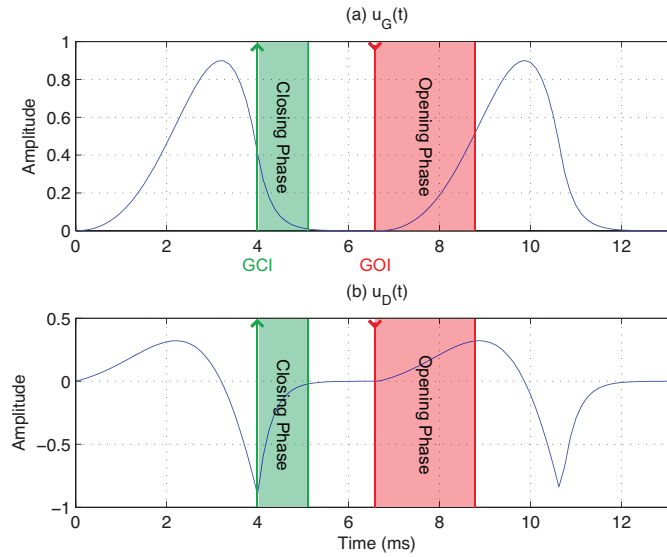


Figure 2.8: LF model of (a) $u_G(n)$ and (b) $u_D(n)$ with glottal phases overlaid.

cannot make this assumption if they are to self-oscillate. The final assumption concerns the validity of an all-pole model for the vocal tract for nasalised vowels. Nasal vowels are distinct oral vowel forms in certain European languages, such as in the French *bon* /bõ/. Nasalised vowels are produced when the velum is lowered to allow air to pass into both the oral and nasal tract, introducing zeros into the vocal tract transfer function [18]. The use of an ARMA to model both poles and zeros in the vocal tract is reviewed in Section 2.6.2 which allows for nasalised vowels, though they are restrictive in that they require a parametric model of the glottal excitation signal. Throughout this work it is assumed that these assumptions are valid for glottal-synchronous speech processing.

2.5 Glottal Waveform Models

Models of the vocal tract transfer function are well-rooted in the theory of acoustics and Digital Signal Processing (DSP) for which many quantization schemes have been developed to reduce the number of parameters that represent them [22]. Two cycles of glottal volume velocity $u_G(n)$ and voice source $u_D(n)$ are shown in Figure 2.8 with the LF model (Section 2.5.1). Models of the glottal waveform take similar forms: a large discontinuity occurs in $u_D(n)$ at the GCI, followed by a closing phase, closed phase, GOI

and opening phase. The lengths of closing and opening phase are often in practice much shorter than depicted, particularly the closing phase whose duration is often within one sampling period.

Among the simplest models of glottal volume velocity is the ‘two-pole’ model which consists of two real poles close to unity [23, 24],

$$G(z) = \frac{K_G}{(1 - z_1 z^{-1})(1 - z_2 z^{-1})} \quad (2.19)$$

where K_G is a gain factor and z_1 and z_2 are poles. The joint transfer function of the two-pole model, where $z_1 = 1 - \epsilon$, and the lip radiation filter $R(z)$ is

$$G(z)R(z) = \frac{K_G K_R (1 - z^{-1})}{(1 - z_1 z^{-1})(1 - z_2 z^{-1})} \simeq \frac{K_G K_R}{(1 - z_2 z^{-1})}. \quad (2.20)$$

In applications where the vocal tract is estimated from speech signals, it is necessary to remove the second pole so as to remove the influence of the glottal excitation. This preemphasis of the speech signal is applied for this purpose in Section 2.6.1. Although the two-pole assumption is widely used, it does not model the time-domain waveform of the glottal volume velocity well so higher-order all-pole filters have been proposed [25]. A number of alternative models exist that fall into three broad categories: curve fitting, physical modelling and error minimization, whose uses are dependent upon the type of application. These models are reviewed here.

2.5.1 Curve Fitting Models

Curve fitting models employ piecewise-continuous time-domain models to approximate glottal waveforms. They are conceptually and computationally straightforward and are widely used in speech synthesis [26] but estimating their parameters for analysis is a more challenging task [27].

The Rosenberg model [28] proposes a three-segment model of glottal volume velocity, $u_{Rosenberg}(n)$, that varies between a triangular and trapezoidal shape in six steps: 1 triangular, 1 polynomial, 3 trigonometric and 1 trapezoidal. The first trigonometric curve

is given by

$$u_D^{Rosenberg}(n) = \begin{cases} \frac{1}{2}(1 - \cos(\pi n/N_1)) & \text{for } 0 \leq n < N_1, \\ \cos(\frac{\pi}{2}(n - N_1)/N_2) & \text{for } N_1 \leq n < N_1 + N_2, \\ 0 & \text{for } N_1 + N_2 < n \leq N - 1. \end{cases} \quad (2.21)$$

where N_1 and N_2 are modelling parameters. Fant proposed a series of models for the derivative of glottal volume velocity, $u_D(n)$ including two piecewise trigonometric functions [29,30] and the popular seven-parameter Liljencrants-Fant model [31] which includes an exponential recovery phase,

$$u_D^{LF}(n) = \begin{cases} 0 & \text{for } 0 \leq n < N_o \\ E_0 e^{\alpha(n-N_o)} \sin(\omega_0(n - N_o)) & \text{for } N_o \leq n < N_e \\ -E_1(e^{-\beta(n-N_e)} - e^{-\beta(N_c-N_e)}) & \end{cases} \quad (2.22)$$

where $E_0 = E_e/e^{\alpha(N_e-N_o)} \sin \omega_0(N_e - N_o)$, $E_1 = E_e/(1 - e^{-\beta(N_c-N_e)})$ and E_e , α , β , N_o , N_e , N_c and ω_0 are model parameters. The dependence of E_e on E_0 and E_1 constrains $u_D^{LF}(n)$ to be piecewise-continuous. The glottal volume velocity can be found by analytical integration of the formulation in (2.22). Variants on the LF model exist to model the glottal waveform in more detail, modelling skewness [32] and nonlinearities caused by coupling between the glottis and vocal tract [33]. Procedures for estimating parameters have also been proposed [27], [32]. Other piecewise models include KLGLOTT88 [34], Fujisaki and Ljungqvist [35] and Alku and Backstrom [36]. While providing a convenient mathematical description, curve fitting models impose characteristics that do not necessarily reflect physical changes in the vocal folds.

2.5.2 Physical Modelling

Physical modelling describes the glottis as one or more damped, coupled systems of masses. Though well-rooted in classical physics, they often contain a very large parameter set that is difficult to estimate. Various one-mass models have been proposed [37,38] which

model a mass moving perpendicular and parallel to the flow of air respectively. The models produce perceptually good results when used to generate glottal volume velocity waveforms for speech synthesis but often fail to produce the physiological detail observed with invasive measurements [39]. Multiple-mass and body cover models [39–41] have been shown to model the glottal volume velocity with much greater detail. The assumption of linear separability between the glottal excitation and vocal tract made in Section 2.4 does not hold with most physical models as most cannot sustain oscillation without considering nonlinear interaction between them [39]. Integration of physical models into existing vocal tract models is therefore problematic as the assumption of linear separability is often made.

2.5.3 Error-Minimizing Models

An *error-minimizing* model is one which is not motivated by curve fitting or physical modelling but produces accurate reproduction of the glottal excitation signal as the sum of one or more empirically-derived signals. The approach is commonly seen in the analysis-synthesis and compression of speech where a compact representation that minimises resynthesis error is of greater importance than a physiological derivation.

As a first approximation to the voice source, an impulse train at the GCIs can be used [23]. An extension is the *multi-pulse* model that minimizes synthesis error with an excitation signal comprising of delayed and summed impulses [42] that have no physiological connection but help to minimise the mean square error of the synthesised speech. The Code-Excited Linear Prediction (CELP) CODEC [43] contains a stochastic codebook of noise-like excitation signals. A *long-term predictor* adds a codebook entry and a delayed excitation, chosen such that the mean-square error between the synthesised and original excitation is minimised. The optimal codebook delay is likely to be equal to the period of voicing but may not be if the error is minimised with a different value. A combination of noise and glottal excitation codebooks has been devised where glottal excitation codebooks contain shifted versions of *prototype* excitations [44]. This two-codebook technique has more recently been applied to wideband speech coding for telephony [45, 46].

A key objective of this thesis is the investigation into data-driven glottal excitation

modelling – a type of error-minimizing model – where machine learning techniques are applied to speech databases that have been inverse-filtered and segmented into glottal cycles. Processing glottal excitation in this way is made possible by the GCI detection techniques presented in Chapter 4. The aim is to decompose the glottal cycle into a set of basis functions that provide an accurate and compact representation whose parameters are straightforward to estimate. Prototype glottal flow waveforms, such as those used in glottal pulse codebooks, can then be determined by clustering decomposition spectra.

2.6 Estimating Vocal Tract Transfer Function and Glottal Waveforms

In many speech processing applications it is advantageous to estimate the transfer function of the vocal tract from speech signals. This allows the estimation of the glottal source signals [1, 47, 48] by inverse-filtering for use in coding [44], modelling [31] and analysis of pathological speech [3]. Early efforts for noninvasively determining the vocal tract formants involved tuning an LCR filter [47, 49]. Automatic methods include nonstationary linear prediction [50]: a class of numerical and computationally-efficient techniques for estimating the vocal tract AR parameters from speech signals as a function of time. Iterative [51], nonlinear [52] and cepstrum processing techniques [53] have also been developed for the estimation of the vocal tract filter and excitation signals.

2.6.1 Linear Predictive Coding (LPC)

The all-pole model of the vocal tract defines voiced speech as a delayed and scaled sum of previous samples added to the glottal excitation signal. Assuming vocal tract resonances have high gain, the second term in (2.18) dominates,

$$s(n) \simeq \sum_{j=1}^p a_j s(n-j), \quad (2.23)$$

therefore the dependence on $u_D(n)$ is removed. The linear prediction residual, $e(n)$, is the difference between the speech signal and its autoregressive description [23]:

$$e(n) = s(n) - \sum_{j=1}^p a_j s(n-j) = s(n) - a_1 s(n-1) - a_2 s(n-2) - \cdots - a_p s(n-p) \quad (2.24)$$

or in the z -domain, $E(z) = S(z)A(z)$. The optimal a_k minimise the square error,

$$Q_E = \sum_{n \in \{F\}} e^2(n), \quad (2.25)$$

where $\{F\}$ is a frame of windowed speech and Q_E is cumulative square error. It can be shown [9] that the problem can be represented as the *normal* equations,

$$\sum_{j=1}^p \phi_{i,j} a_j = \phi_{i,0} \text{ where } \phi_{i,j} = \sum_{n \in \{F\}} s(n-i)s(n-j). \quad (2.26)$$

Represented in matrix form,

$$\Phi \mathbf{a} = \mathbf{c} \Rightarrow \mathbf{a} = \Phi^{-1} \mathbf{c}. \quad (2.27)$$

The nature of Φ is determined by the choice of frame. Windowed frames of infinite extent are termed Autocorrelation LPC [24]; finite non-windowed frames during the glottal closed phase are termed Closed-Phase Covariance LPC [2, 54], defined as follows.

Autocorrelation LPC

Let frame $\{F\}$ be an infinite frame of windowed speech such that

$$\phi_{i,j} = \sum_{n=-\infty}^{+\infty} \tilde{s}(n-i)\tilde{s}(n-j) \quad (2.28)$$

where $\tilde{s} = w(n)s(n)$ and $w(n)$ is a window of typically 20 – 30 ms with 50 % overlap. The infinite sum leads to symmetrical, positive-definite and Toeplitz matrix Φ where $\phi_{i,j} = \phi_{|i-j|,0} = R_{|i-j|}$ and R_k is an autocorrelation function at lag k . The inverse Φ^{-1} can be solved using the computationally-efficient *Yule Walker* equations [50], scaling $O(p^2)$. The resulting filter is also guaranteed stable [24].

A problem arises when the spectrum of the glottal excitation signal, $u_D(n)$, is considered. Although the glottal excitation signal is ignored in (2.23), linear prediction nevertheless remains a spectral estimation technique and cannot distinguish those spectral components arising from the excitation signal and those from the vocal tract. The two-pole model of the glottal volume velocity was discussed in Section 2.5, where it was shown that lip radiation cancels one of the poles, leaving a second pole near unity. A preemphasis filter $P(z)$ cancels the remaining pole,

$$P(z) = 1 - \mu z^{-1} \quad (2.29)$$

where $\mu \sim 1$. It is often chosen in the range $0.9 \leq \mu \leq 1.0$, although it is relatively insensitive to exact values. It is suggested in [55] that an ‘optimal’ value of μ is given by

$$\mu = \frac{R_1}{R_0}. \quad (2.30)$$

The excitation source in unvoiced frames is close to being spectrally white, therefore preemphasis is unnecessary. A discussion of the effectiveness of $P(z)$ in removing the spectral contribution of $u_D(n)$ can be found in Section 5.5.1, where it is suggested that a more effective filter can be achieved by convolution with a least-squares inverse *prototype* glottal pulse. This choice of time-invariant preemphasis filter is designed to provide improved spectral whitening of the voice source compared with the first-order filter in (2.29). During the glottal cycle, a period of time exists during which the glottis is closed and does not excite the vocal tract. The concept of closed-phase LPC aims to exploit this by constraining the analysis window to the closed phase, relinquishing the need for preemphasis entirely and potentially finding a more accurate model of the vocal tract.

Closed-Phase Covariance LPC

Consider a finite segment of speech such that [56]

$$\phi_{i,j} = \sum_{n=0}^{N-1} \tilde{s}(n-i)\tilde{s}(n-j) \quad (2.31)$$

where $\tilde{s}(n) = s(n)$, $n_r^c \leq n < n_r^o$, $N = n_r^o - n_r^c + 1$ and n_r^c and n_r^o denote the beginning and ending of the closed phase for cycle r respectively. No windowing is required, giving infinite spectral resolution. Much shorter analysis windows of > 2 ms can be used. A disadvantage lies in computational complexity as Φ is symmetric but not Toeplitz, scaling approximately $O(p^3)$. The resulting AR coefficients are also not guaranteed stable [24]. The stability issue can be improved by multi-glottal closed phase analysis [57] that includes adjacent glottal closed phases $\mathcal{C}_n$ in the calculation of the covariance matrix Φ .

Detection of the glottal closed phases has been made from X-ray video recordings [47], laryngograph [56, 57] and with automatic estimation from the speech waveform [33, 54]. By considering the energy in the closed phase following inverse-filtering, which is known to be near-zero when the glottis is closed, closed-phase covariance LPC can be shown to produce better results than fixed-frame autocorrelation LPC [24] but at the expense of requiring knowledge of the glottal closed phases and the risk of producing unstable filters. Filter instability is not problematic for inverse-filtering as $1/V(z)$ is an all-zero filter. In the case of resynthesis a stable vocal tract filter is required.

Framing errors in the closed phase increase the likelihood of instability if glottal excitation is included in the analysis; this is particularly problematic for those, particularly female, voices where the closed phase does not exist [9]. Fixed-frame covariance LPC can produce improved results compared with autocorrelation LPC but with increased computational cost [9]. For these reasons autocorrelation LPC is used throughout this work, in part with enhanced preemphasis filtering as discussed in Chapter 5.

LPC Voice Source Estimation

The assumption that glottal waveforms are filtered by a linearly-separable AR filter allows the cancelling, or *inverse-filtering*, of the vocal tract poles using an MA filter with identical roots [1, 47, 48],

$$\hat{U}_D(z) = \hat{U}_G(z)R(z) \simeq \frac{S(z)}{\hat{V}(z)}, \quad (2.32)$$

where $\hat{\cdot}$ denotes estimation. The acoustic delay of the vocal tract, $z^{-p/2}$ and the propagation delay from the lips to the microphone are ignored such that $u_D(n)$ and $s(z)$ are

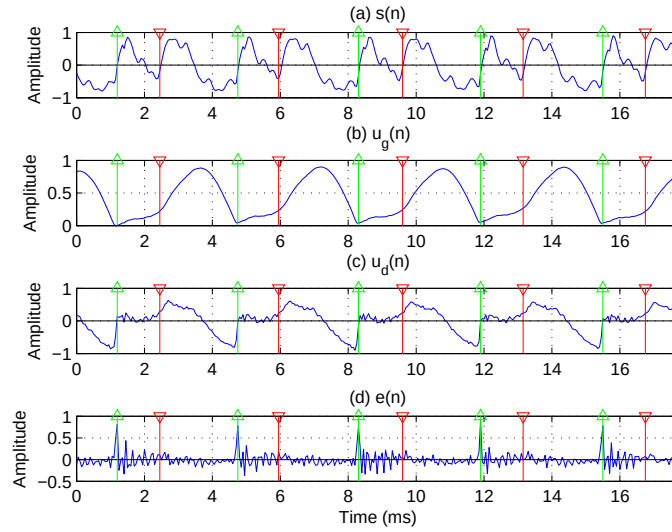


Figure 2.9: (a) Speech signal $s(n)$, (b) Glottal volume velocity $u_g(n)$, (c) Glottal pressure $u_d(n)$ and (d) LPC residual $e(n)$. GCIs are overlaid as green $\triangle$ and GOIs as red ∇ .

time-aligned. An inverse lip radiation filter, $R^{-1}(z)$, may be applied if the glottal volume velocity $\hat{U}_G(z)$ is required. Inversion of the model $R(z) = 1 - z^{-1}$ is not possible as the initial conditions of the inverting integrator are not known, so instead a ‘leaky integrator’ is applied, whose pole is placed close to the unit circle [58],

$$\frac{1}{\hat{R}(z)} = \frac{K_R^{-1}}{1 - (1 - \epsilon)z^{-1}}. \quad (2.33)$$

The glottal volume velocity can then be estimated in a similar manner to (2.32),

$$U_G(z) = \frac{S(z)}{\hat{R}(z)\hat{V}(z)}. \quad (2.34)$$

Let the LPC residual be the inverse-filtered preemphasised speech,

$$E(z) = \frac{P(z)S(z)}{\hat{V}(z)}. \quad (2.35)$$

It is approximately the first derivative of $u_d(n)$ and the second derivative of $u_g(n)$. It is maximally white, and is useful in coding with noise codebooks and GCI detection, as discussed in Sections 2.5.3 and 2.8 respectively.

Figure 2.9 shows a few cycles of $s(n)$, $u_G(n)$, $u_D(n)$ and $e(n)$ with GCIs ($\triangle$) and GOIs (∇) overlaid. All display instants of discontinuity at the GCIs, in particular in $e(n)$ which displays impulsive features. The flat closed phase is visible on all three glottal waveforms, although glottal opening instants are difficult to find in $e(n)$.

2.6.2 Simultaneous Estimation of Vocal Tract and Glottal Excitation

The error function in (2.24) assumes the excitation signal is negligible. However, nonzero glottal excitation can be used by applying parameteric models discussed in Section 2.5, giving two error signals that can be minimised with respect to both vocal tract and voice source parameters [59]. The first error signal is the difference between the parameterized glottal volume velocity, $G(z)$, and the estimated glottal volume velocity, $\hat{U}_G(z)$,

$$E_1(z) = G(z)R(z) - \hat{U}_G(z)R(z) = G(z)R(z) - \frac{S(z)}{\hat{V}(z)}. \quad (2.36)$$

The second error signal is the difference between the true and synthesised speech signals,

$$E_2(z) = S(z) - \hat{S}(z) = S(z) - G(z)\hat{V}(z)R(z). \quad (2.37)$$

The glottal excitation signal used in [59] is the two-pole model, whose parameters are the two polynomial coefficients and their respective weights. The model must be correctly aligned with $\hat{U}_G(z)$ for which GCIs are required. An alternative approach is the ARMA model of speech production [35], given by

$$A(z)S(z) = B(z)G(z), \quad (2.38)$$

where $B(z)$ and $A(z)$ are the numerator and denominator of the ARMA model respectively.

The corresponding error function is given by

$$E(z) = S(z) - \frac{B(z)}{A(z)}U(z). \quad (2.39)$$

Where $U(z)$ is a driving function. The notation differs from that used elsewhere in this work; $A(z)$ corresponds to $V(z)$ but $B(z)$ and $U(z)$ share a combination of $R(z)$, $G(z)$ and $U_G(z)$. This is a nonlinear minimization problem that can be simplified by minimizing $A(z)E(z)$ and estimating the coefficients for $A(z)$ and $B(z)$ [35].

Various glottal excitation models have been applied to the simultaneous estimation of vocal tract excitation parameters [60,61]. Models applying Autoregressive Exogenous Input (ARX) have also been used [62,63]. These approaches all require an existing parametric model of glottal excitation, although no parametric model is capable of synthesizing all the waveforms produced by the glottis as discussed in Section 2.5. They must also be correctly time-aligned, requiring knowledge of the GCIs which cannot always be reliably recovered from recorded speech signals.

2.6.3 Homomorphic Processing

The complex cepstrum can be used in the deconvolution of signals as it is a transform that turns a convolution into sum. This *homomorphic* processing allows the separation of a speech signal into glottal excitation and vocal tract components. Let $S(\omega) = \mathcal{F}\{s(n)\} = ||S(\omega)||e^{j\Theta(\omega)}$ where $\mathcal{F}$ is the Discrete-Time Fourier Transform (DTFT). The filtering of $U_D(\omega)$ with $V(\omega)$ is achieved by multiplication in the frequency domain, $S(\omega) = U_D(\omega)V(\omega)$. Multiplication is transformed into addition by taking logs,

$$C(\omega) = \log ||S(\omega)||^2 = \log ||U_D(\omega)||^2 + \log ||V(\omega)||^2. \quad (2.40)$$

The inverse DTFT provides a time-domain signal termed the complex cepstrum [64],

$$c(n) = \mathcal{F}^{-1} \{ \log ||U_D(\omega)||^2 \} + \mathcal{F}^{-1} \{ \log ||V(\omega)||^2 \}. \quad (2.41)$$

The additive representation of the glottal excitation and vocal tract permits their separation by removing the relevant cepstral coefficients $c(n)$. It is assumed that the vocal tract resonances cause slow changes in the spectrum and correspond to low order cepstral coefficients; glottal pulses cause quick changes in the spectrum and correspond to high

order coefficients [64]. Define a rectangular window,

$$l(n) = \begin{cases} 1 & \text{for } n_0 \leq n \leq N - n_0 \\ 0 & \text{for } 0 \leq n \leq n_0, N - n_0 < n < N \end{cases} \quad (2.42)$$

where n_0 is empirically chosen as $0.002f_s$ [53] and N is the length of the DTFT. Windowing $c(n)$ is the process of liftering (an anagram of filtering) which removes components according to the rectangular window in (2.42),

$$\tilde{c}(n) = c(n)l(n). \quad (2.43)$$

Recovering the approximation to $u_D(n)$, $\hat{u}_D(n)$, involves the inverse cepstrum of $\tilde{c}(n)$ with the phase from $s(n)$, $\Theta(\omega)$,

$$\hat{u}_D(n) = \Re \left\{ \mathcal{F}^{-1} \left[|\tilde{U}_D(\omega)| e^{j\Theta(\omega)} \right] \right\} \quad (2.44)$$

$$\text{where } \tilde{U}_D(\omega) = 10^{\mathcal{F}\{\tilde{c}(n)\}}. \quad (2.45)$$

The ability of liftering to deconvolve two signals depends upon the degree to which their components are separated in the cepstral domain; overlapping (cross-product) components render perfect deconvolution impossible [64]. The correct choice of n_0 is also important. The complex cepstrum is often used in the blind deconvolution of reverberant speech signals [65–67] whose effects are reviewed in Section 2.9.1.

2.7 Measuring Glottal Waveforms (Invasive Methods)

Invasive measurements can provide high quality signals by circumventing unwanted signals such as acoustic noise, reverberation and spectral filtering by the vocal tract. The disadvantage compared with noninvasive measurements is a requirement for specialised equipment that cannot be used remotely. The following is a review of invasive methods commonly used in speech science.

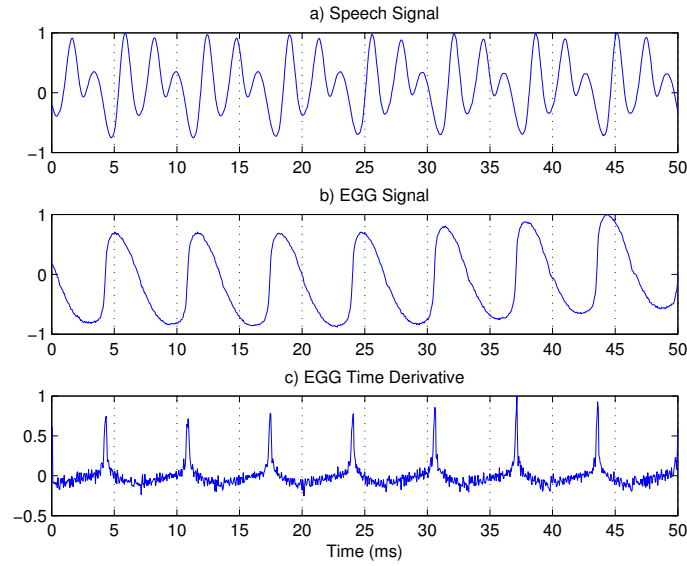


Figure 2.10: (a) Speech signal and (b) Time-aligned EGG signal (c) Time derivative of EGG signal.

2.7.1 EGG

The Electroglottogram (EGG) signal, $\eta(n)$, [7] is a measurement of the electrical conductance of the glottis, usually captured contemporaneously with speech recordings. The measured signal is proportional to the glottal contact area, whose time derivative (DEGG) during voiced speech contains short, high-amplitude impulsive temporal features due to glottal closure and smaller features of opposite sign due to glottal opening. An example of a voiced speech segment, the corresponding EGG recording and its time derivative is shown in Figure 2.10. Time alignment between the speech and EGG is achieved by fixing a talker's mouth a known distance from a microphone and subtracting the delay. The EGG waveform is dissimilar to the voice source signal as the former estimates contact area and the latter estimates a pressure waveform which are not generally proportional [68]. Section 4.4.4 discusses some of the differences between information provided by each signal.

The advantage of EGG signals is that, compared with speech signals, they are relatively simple to analyse. They are particularly useful for the estimation of GCIs and GOIs as a reference for speech-based algorithms [69]. Chapter 3 reviews some existing EGG-based GCI/GOI detection algorithms and proposes a novel and robust technique.

2.7.2 Pneumotachograph

The purpose of Pneumotachograph is to measure the volume velocity of air through mask placed in front of the mouth, U_m [70]. The mask contains a fine stainless steel wire screen that subjects the airflow to a known acoustic resistance, Z_m , and a differential pressure transducer measures the pressure drop, P_m , across this resistance. The volume velocity can then be estimated with the acoustic equivalent of Ohm's law,

$$U_m = \frac{P_m}{Z_m}. \quad (2.46)$$

Practical measurements show that this device is limited because it is acoustically active, both distorting the radiated speech and limiting the high-frequency response because of the transfer function between the mouth radiation and front of the screen [71]. The Rothenberg Mask [71] addresses these problems with an optimized circumferentially vented mask, whose response is consistent up to ~ 1 kHz [72].

2.7.3 Laryngoscopy

Laryngoscopy encompasses many techniques for obtaining a view of the glottis. The contemporary approach involves inserting a fireoptic camera to view the glottis without significantly impairing the patient's ability to speak or sing. Despite providing a good view of the glottis, the frame rate of conventional videoendoscopy cameras is insufficient to capture the fast motion of the glottis. Instead stroboscopy [73] with a strobe light and high shutter speed are used that can be interpreted by an *Otolaryngologist*, albeit temporally aliased. It is also suitable only for periodic vibration.

Recent advances in videoendoscopy employ a high-speed camera, capturing 6000–10000 frames per second [74]. This enables accurate estimation of glottal abduction/adduction, with further application in digital kymography for examining glottal adduction/abduction characteristics as a function of time [75]. Figure 2.11 shows a frame of high-speed laryngoscopy with an open glottis. The zipper-like motion of the glottis can be seen as viewed normal to the page, where the furthestmost fold is slightly more adducted

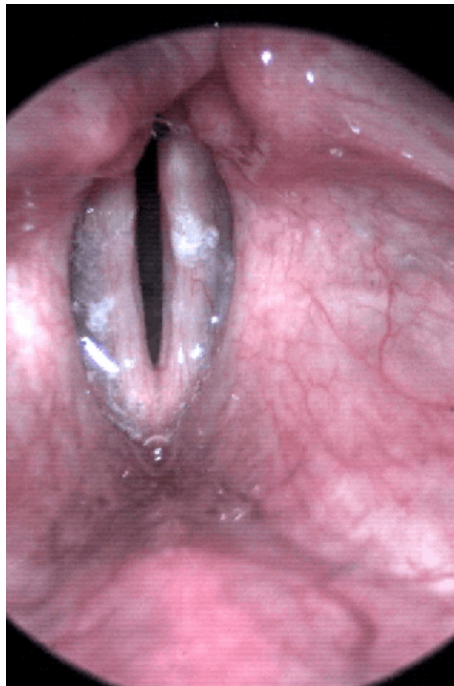


Figure 2.11: Frame of high speed laryngoscopy showing an open glottis.¹

than the near fold. This phenomenon often observed with endoscopic measurements of both normal and pathological voice [76].

¹Data provided by D. D. Mehta, SM, Speech & Hearing Bioscience & Technology Program, Harvard-MIT Division of Health Sciences & Technology. Recording performed at the Center for Laryngeal Surgery & Voice Rehabilitation, Massachusetts General Hospital (Co-Directors: S. M. Zeitels, MD, & R. E. Hillman, PhD).

2.8 GCI/GOI Detection

The remainder of this chapter discusses practical applications of speech processing, beginning with GCI/GOI detection, followed by application context in Section 2.9, objective quality measures in Section 2.10 and test corpora in Section 2.11. Central to glottal-synchronous speech processing is the detection of periodicity in voiced speech. During the glottal cycle, energy is imparted into the vocal tract at the glottal closure instant which causes a discontinuity in the voice source signal $u_D(n)$. A second, usually weaker, discontinuity is observed at the glottal opening instant. The GCI is usually chosen to delimit glottal cycles it forms the most prominent feature of the glottal cycle. A brief overview of the methods for GCI/GOI detection is given here with full review in Chapter 4.

2.8.1 Noninvasive Methods

Existing noninvasive methods for GCI detection from speech signals do so in three stages: (a) Preprocessing, to produce a signal containing features at the GCIs that can be easily detected, (b) Event detection, to identify GCI/GOIs from the output of (a), and (c) Postprocessing, to remove incorrect detections from (b) by applying suitable heuristics. The DYPSA algorithm [77] adheres closely to this model. The preprocessor finds the linear prediction residual, $e(n)$, by estimating the vocal tract filter $V(z)$ and using it inverse-filter the speech signal $e(n)$. The residual signal is approximately a train of impulses at the GCIs with an additive noise component from modelling errors and acoustic noise sources as shown in Figure 2.9(d). Event detection finds a set of GCI *candidates* by applying the group delay function [78] and searching for negative-going zero crossings which occur at the centre of gravity of each impulse. A number of false detections occur due to the presence of additive noise. The postprocessor is an N -best dynamic programming algorithm [79] that finds the best path through the candidates according to a set of costs.

Detection of glottal opening instants is a more challenging task as they are often very low in amplitude compared with GCIs. An algorithm is presented in Section 4.4 that achieves good GCI and GOI detection from the voice source signal, $u_D(n)$.

2.8.2 Invasive Methods

The EGG signal, shown in Figure 2.10, is more straightforward to analyse than the speech signal as it is a direct measure of the glottal contact area that is free from distortion by the vocal tract and acoustic noise. Analysis of EGG is less prone to detection errors so the postprocessing stage required for noninvasive GCI/GOI detection is less complicated [69]. A detailed analysis of the EGG signal and existing methods for detection of GCIs and GOIs is given in Chapter 3. Laryngoscopy provides another estimation of glottal contact area, achieved by counting the number of adjacent dark pixels in the laryngoscope image. The estimation is very accurate as recorded images are of sufficiently high resolution to reduce ambiguity.

The assumption that glottal contact area and glottal airflow provide the same glottal closure and opening instants is challenged in Section 4.4.4 where it is shown that discrepancies can often occur. There is also no universally-accepted definition of the GOI [80] as it often occurs as a finite phase and not a single instants in time. The exact definition is heavily dependent upon the application as discussed in Chapter 4.

2.8.3 Voiced/Unvoiced/Silence Detection

Glottal-Synchronous techniques are applicable to voiced and mixed speech only. However, much information is conveyed by unvoiced speech where glottal-synchronous techniques can fail and may even be detrimental to intelligibility. Voiced/Unvoiced/Silence (VUS) detection is necessary to ensure that different segments of speech are treated correctly and is used extensively in the applications presented in Chapter 6.

The following is a description of the VUS detector employed in this work that applies the *Atal* feature set, described in [81], which further shows that the features can be modelled as a multivariate Gaussian distribution. An extension is described in [82] which augments the feature set with delta and delta-delta features to improve accuracy [83]. An additional untrained decision algorithm for situations when no training data is available is described in Section 6.2.2.

VUS Features

In order to determine the VUS classification, the following features are computed based upon 20 ms frames of speech.

1. Zero crossing rate indicates how the energy of the speech signal is concentrated in frequency; voiced segments of speech have a low zero crossing rate. This measure varies significantly for silent periods because of ambient noise.
2. Normalized energy is an approximate measure of voicing and is greatest during voiced segments of speech.
3. Normalized autocorrelation coefficient at one sample delay is a strong indicator of the spectral whiteness of unvoiced speech whose value is close to zero. The periodicity of voiced speech gives a value close to one.
4. Mean spectral slope, estimated by the first covariance LPC coefficient, is steeper for voiced speech owing to the spectrum of $u_D(n)$.
5. Energy in LP residual is a good indicator of the strength of formants present in voiced segments of the speech signal. The five coefficients are then augmented with their delta and delta-delta coefficients to produce a 15 dimensional feature vector $\mathbf{x}_i$ for frame i [83].

Trained VUS Detection

If training data is available, each class $\omega \in \{\mathcal{V}, \mathcal{U}, \mathcal{S}\}$ is modelled by a multivariate full covariance Gaussian distribution [81], whose parameters are derived from labelled training data. The maximum likelihood estimate of the mean vector $\mathbf{m}_\omega$ and the covariance matrix Σ_ω is determined for each class and the relative frequency of each class is used to determine the prior probabilities $P\{\omega\}$. The probability of a feature vector $\mathbf{x}_i$ belonging to class ω is determined using Bayes's rule,

$$P\{\omega|\mathbf{x}_i\} = \frac{P\{\omega\}f_X(\mathbf{x}_i|\omega)}{f_X(\mathbf{x}_i)} \quad (2.47)$$

where $f_X(\mathbf{x}_i|\omega)$ is the class likelihood (determined by $\mathbf{m}_\omega$ and Σ_ω) and the total likelihood is estimated using

$$f_X(\mathbf{x}_i) = \sum_{\omega} P\{\omega\} f_X(\mathbf{x}_i|\omega). \quad (2.48)$$

The classification of frame i may only depend on $\mathbf{x}_i$ in which case it is sufficient to use the numerator of (2.47), where the class is determined as $\max_{\omega} P\{\omega\} f_X(\mathbf{x}_i|\omega)$. When classification is a time-dependent process it is necessary to base the decision on $P\{\omega|\mathbf{x}_i\}$. Examples of this trained VUS detection can be found in [82] and [84].

Alternative VUS

A number of other measures exist that are not considered here, such as the Ink Measure [85] which provides similar results to zero crossing rate. For the purposes of this thesis the described method is sufficient, although VUS detection and the closely-related Voice Activity Detection (VAD) are areas of ongoing research. An untrained glottal-synchronous waveform similarity measure is used for VUS detection in Chapter 4 which is particularly efficient as it is implemented into the dynamic programming postprocessor for a GCI/GOI detector, sharing many of the features already derived. The method proposed in this section is useful as a ‘bolt-on’ VUS detector that requires few parameters.

2.8.4 Pitch Detection

Similar to the detection of GCIs/GOIs is the concept of pitch detection, which has played an important role in speech synthesis, recognition and as metadata in multimedia applications [86]. It is also used in speech quality assessment algorithms such as ITU-T P.563 [87] and Low-Complexity Quality Assessment (LCQA) [88]. Pitch detection differs from GCI/GOI detection in that the pitch only need be known within any given frame, such that the exact instant of a particular event is not required. A GCI/GOI detection algorithm may therefore be used as a pitch estimator. Various approaches exist whose approaches are similar to those found in GCI/GOI detection, including the cross-correlation based RAPT algorithm [89], the autocorrelation-based pitch detector in ITU-T P.563, and difference function-based YIN [90].

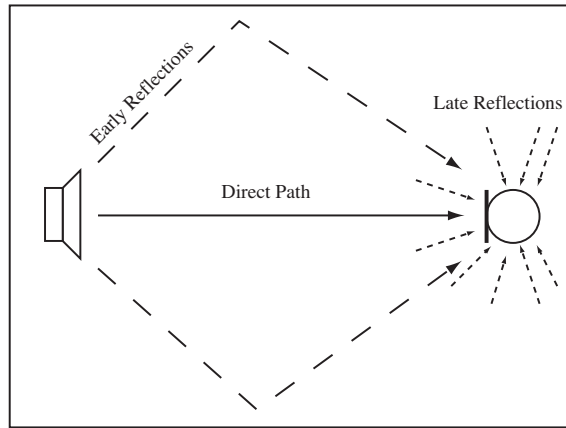


Figure 2.12: Process of reverberation.

2.9 Application Context

Glottal-synchronous speech processing lends itself to a number of existing applications in speech processing, explored in Chapter 6. Additional glottal-synchronous applications not reviewed in this thesis include speech synthesis [91] and pathological speech analysis that relies upon knowledge of the open quotient (OQ) [8].

2.9.1 Dereverberation

Consider a speech signal, $s(n)$, produced in a reverberant room and observed by an M -element microphone array positioned at a distance from the source. In many modern telecommunication applications, speech signals are obtained in enclosed spaces such as office rooms, with the talker situated at a distance from the microphone. Each microphone receives the direct-path speech signal and a number of reflected signals from walls and hard objects as shown in Figure 2.12, reducing intelligibility and perceived speech quality [92]. The aim of dereverberation is to suppress the reflected signals and reinforce the direct-path speech signal. The observation at microphone m is

$$x_m(n) = h_m(n) * s(n), \quad m = 1, 2, \dots, M \quad (2.49)$$

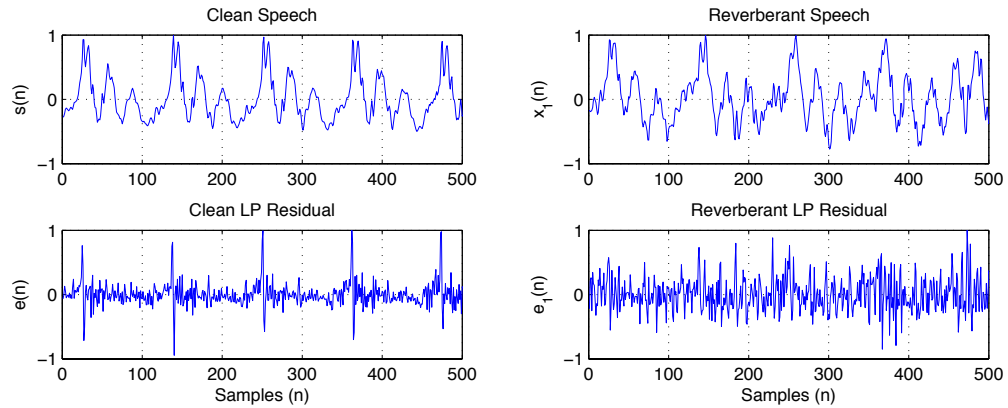


Figure 2.13: Effects of reverberation on LPC residuals: (left) clean speech and clean residual, (right) reverberant speech and reverberant residual.

where h_m is the L -tap impulse response of the acoustic channel between the source to the m th microphone. It has been demonstrated that reverberation mainly affects the LP residual. Studies on the effect of reverberation on voiced speech LP residuals [93, 94] have further shown that the room impulse response results in additional spurious peaks of similar amplitude to the original excitation peaks, as demonstrated in Figure 2.13 which shows an example of clean and reverberant speech and their corresponding LPC residuals. It has been observed that in multiple time-aligned observations from a beamformer, the peaks due to GCIs are correlated, while those due to reverberation are not [95]. This is due to the differing path lengths of the reflected speech to each microphone, which may involve numerous reflections from multiple surfaces. It has been further observed that reverberation components are uncorrelated between neighbouring glottal cycles. This has motivated the development of several speech dereverberation algorithms [93, 95, 96] which reduce the effects of reverberation by attenuating such uncorrelated components.

A multichannel dereverberation algorithm is presented in Section 6.1 where the periodic temporal nature of speech and the spatial nature of reverberant speech are exploited in a process termed spatiotemporal averaging. A multichannel GCI detection algorithm, described in Section 4.5.3, is first applied to detect GCIs from the reverberant speech. A Delay-and-Sum Beamformer (DSB) then points a beam of sensitivity in the direction of the speaker; the algorithm then averages neighbouring glottal cycles from the DSB signal,

suppressing the unwanted signals and reinforcing the wanted speech signal. Delay-and-sum beamforming is common to both GCI detection and dereverberation and is described in the following subsection.

Results using objective measures demonstrate that a measurable improvement in speech quality can be achieved with spatiotemporal averaging. Perceptually, reverberation components are attenuated and the microphone sounds closer to the talker.

Delay-and-Sum Beamforming

Speech processing techniques are generally limited to monaural (single-channel) and stereo (two-channel) recordings. Multichannel configurations can be beneficial as they exploit the spatial diversity of speech signals. A relatively straightforward way of doing so is to implement a *spatial averaging* with a DSB. The underlying assumption is that the wanted signal, for example a talker in a noisy and/or reverberant room, is incident on all the microphones in the array. Delaying to compensate for the propagation path length and summing reinforces the wanted signal and attenuates noise and reverberation propagating from different directions. Consider the observation, $x_m(n)$, with DTFT $X_m(\omega)$, whose Time Delay of Arrival (TDoA) relative to a reference microphone is τ_m . Time-aligning and summing across channels yields the single-channel signal

$$\bar{x}(n) = \frac{1}{M} \sum_{m=0}^{M-1} x_m(n - \tau_m). \quad (2.50)$$

The τ_m can be estimated by knowledge of the geometry of the array and the relative positions of the source and array. In general, such information is unavailable so it is necessary to estimate TDoA by a cross-correlation between channels. The Generalized Cross-Correlation PHase Transform (GCC-PHAT) [97] is a relatively robust approach for office environments. Let there be a reference channel, $x_{ref}(n)$, with DTFT $X_{ref}(\omega)$, to which the interchannel delays are referred. The delay, $\hat{\tau}_{GCC}$, is determined by maximising the cross-correlation between channels

$$\hat{\tau}_{GCC} = \arg \max_{\tau} R_{x_{ref}x_m}(\tau), \quad (2.51)$$

where

$$R_{x_{ref}x_m}(\tau) = \int_{-\infty}^{\infty} \frac{X_{ref}(\omega)X_m^*(\omega)}{|X_{ref}(\omega)||X_m^*(\omega)|} e^{j\omega\tau} d\omega. \quad (2.52)$$

The beamformer focuses its beam on a single point in space, providing the τ_m are estimated correctly [93].

2.9.2 Time Scale Modification

Speech time scale modification is a process which alters the length of a segment of speech without significantly affecting its pitch or formant structure. Uses include time scale compression for fast scanning of recorded voicemail messages [98] and time scale expansion for improving the intelligibility of fast or degraded speech in forensic applications. A combination of compression and expansion may also find uses in the synchronization of audio to lip movements in motion video.

The pseudo-periodicity of the voiced speech naturally lends itself to time scale modification as complete glottal cycles may be removed or repeated depending upon whether a compression or expansion of signal duration is desired. This is termed Pitch-Synchronous Overlap-Add (PSOLA) [99] and works well providing the periods are accurately known and cycles are concatenated in such a way that pitch periods are faithfully reproduced. Existing approaches for concatenating periods of voiced speech for time scale/pitch modification, specifically time-domain PSOLA (TD-PSOLA), performs well provided a) pitch periods are accurately known and b) high quality time scale (but not pitch) modification is required. Other approaches include sinusoidal-based [100], LP residual-based (LP-PSOLA) [50, 101], waveform similarity-based (WSOLA) [102] and phase vocoders [103], which address the cases when one or more of these constraints are unfeasible, at the cost of added complexity.

Section 6.3 extends PSOLA by not applying modification during unvoiced (U) speech and the transition times (T) at the beginning and ending of voicing [104, 105]. Although more detailed models of the length of UT segments have been made [105–109], little work has been done to optimize these parameters and the studies generally conclude

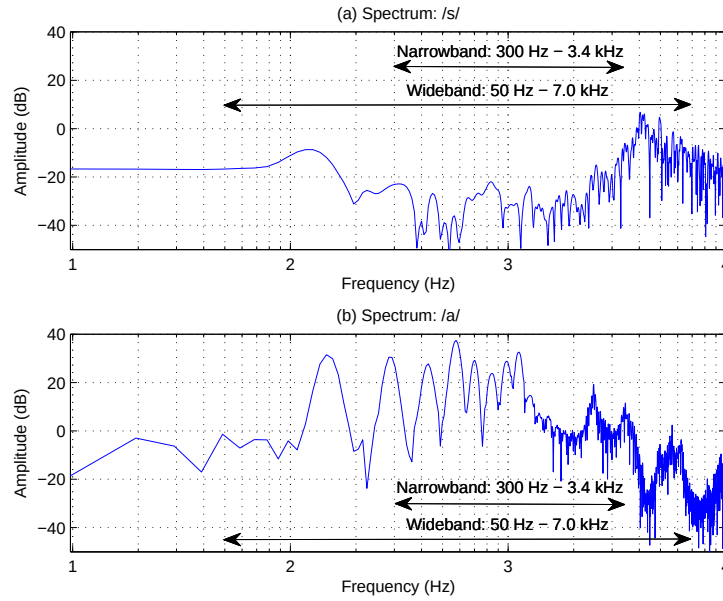


Figure 2.14: Speech Spectra for (a) /s/ (unvoiced) and (b) /a/ (voiced) with narrowband and wideband bandwidths.

that the most perceptually significant artefacts are those arising from the repetition of UT segments. Results with subjective testing show that the proposed technique can provide perceptually superior results than the standard PSOLA, demonstrating the requirement for accurate Voiced/Unvoiced/Silence (VUS) detection in addition to GCIs.

2.9.3 Artificial Bandwidth Extension

Narrowband telephony limits the spectrum of speech to 300 Hz – 3.4 kHz; wideband speech is considered to be 50Hz – 7 kHz. Figure 2.14 shows speech spectra and system audio bandwidths for (a) unvoiced speech missing high-frequency energy, and (b) voiced speech, missing low frequency energy. This bandwidth limitation impairs intelligibility and perceived quality.

The aim of Artificial Bandwidth Extension (ABWE) is to artificially increase the spectral content from narrowband speech signals. Many existing methods apply the source-filter model of speech and estimate spectral envelope of the upper extension band using codebook mapping [110], piece-wise linear mapping [111] and Bayesian methods based on Gaussian Mixture Models (GMMs) [112] or Hidden Markov Models (HMMs) [113]. It

has been demonstrated that for upper extension bands a precise estimate of the spectral envelope is important while the glottal excitation signal extension is less important [5]; straightforward DSP techniques such as spectral translation, mirroring and modulation of the narrowband glottal excitation can give satisfactory results. Relatively little attention has been paid to the ABWE in the lowband where the opposite applies and the correct excitation signal is required. Section 6.4 presents a method that uses a glottal-synchronous data-driven model developed in Chapter 5 to map narrowband excitation signals to their wideband equivalent. Subjective testing shows that perceptually superior results can be achieved when lowband excitation is estimated in this way compared with the existing techniques.

2.10 Speech Quality Measurements

In order to evaluate speech processing algorithms it is often necessary to seek one or more standard measures. Measurement of speech quality can be divided into two categories: objective, where an algorithm provides a numerical measure of quality, and subjective, where many human subjects rate speech quality on a scale which is later averaged.

2.10.1 Objective Measures

Objective measures are applied to speech enhancement in Section 6.2 and resynthesis in Section 5.5.2. They are desirable because they provide fast repeatable measures compared with subjective measures, which can vary significantly unless the time-consuming task of assessing very large numbers of candidates is undertaken. The following quality measurements are often used in speech enhancement and coding [114, 115] and are intrusive, requiring the original and processed signals to quantify the distortion. Two measures are considered here: Segmental Signal-to-Noise Ratio (SSNR) and Bark Spectral Distortion (BSD).

Segmental Signal-to-Noise Ratio

SSNR is a measurement of the mean square deviation between the processed and reference speech, calculated on frames of typically 30 ms, defined as

$$\text{SSNR} = \frac{10}{\mathcal{I}} \sum_{i=0}^{\mathcal{I}-1} \log_{10} \left[\frac{\|\mathbf{s}(i)\|^2}{\|\hat{\mathbf{s}}(i) - \mathbf{s}(i)\|^2} \right] \text{ dB} \quad (2.53)$$

where $\mathbf{s}(i) = [s(iL) \ s(iL+1) \ \dots \ s(iL+L-1)]^T$ is a vector of clean speech signals and $\hat{\mathbf{s}}(i) = [\hat{s}(iL) \ \hat{s}(iL+1) \ \dots \ \hat{s}(iL+L-1)]^T$ is a vector of processed speech signals from frame i , $\mathcal{I}$ is the total number of frames and L is the frame length in samples.

Normalized Bark Spectral Distortion

Normalized BSD is a perceptually-motivated measurement of the deviation of Bark spectra between two signals [114,116], calculated on frames of length ~ 30 ms and overlap ~ 50 %, defined as

$$\text{BSD} = \frac{\sum_{i=0}^{\mathcal{I}-1} \|\mathcal{B}_{\hat{\mathbf{s}}}(i) - \mathcal{B}_{\mathbf{s}}(i)\|^2}{\sum_{i=0}^{\mathcal{I}-1} \|\mathcal{B}_{\mathbf{s}}(i)\|^2} \quad (2.54)$$

where $\mathcal{B}_{\mathbf{s}}(i) = [\mathcal{B}_s(iL) \ \mathcal{B}_s(iL+1) \ \dots \ \mathcal{B}_s(iL+L-1)]^T$ and $\mathcal{B}_{\hat{\mathbf{s}}}(i) = [\mathcal{B}_{\hat{s}}(iL) \ \mathcal{B}_{\hat{s}}(iL+1) \ \dots \ \mathcal{B}_{\hat{s}}(iL+L-1)]^T$ are vectors of Bark spectra for the clean speech $s(n)$ and processed speech signal $\hat{s}(n)$ respectively. The Bark spectrum is calculated in three stages:

1. *Critical band filtering* – The Bark scale has been shown to correlate well with the perception of frequency and is calculated with the following transform [114]

$$b = 13 \tan^{-1} \left(\frac{0.76f}{1000} \right) + 3.5 \tan^{-1} \left(\frac{f}{7500} \right)^2. \quad (2.55)$$

The Bark scale b is then convolved with the transformed transfer function of the critical band filter,

$$10 \log_{10} F(b) = 7 - 7.5(b - 0.215) - 17.5 \sqrt{(0.196 + (b - 0.215)^2)}. \quad (2.56)$$

2. *Perceptual Weighting* – The perception of loudness is a nonlinear function of both

signal amplitude and frequency. The equal loudness contours [4] are an empirically-driven model of loudness perception that represents sounds on the phon scale. Each contour represents the signal amplitude required to maintain a constant perceived loudness as a function of frequency.

3. *Phon-to-Sone Conversion* – A sone is a mapping of the phon scale, defined as the increase in signal power that doubles the subjective loudness [114],

$$\text{Sone} = \begin{cases} 2^{(\text{Phon}-40)/10}, & \text{Phon} \geq 40 \\ (\text{Phon}/40)^{2.642}, & \text{Phon} < 40. \end{cases} \quad (2.57)$$

In practice it is difficult to determine the exact number of phones received in a subject's ear. By assuming that normal listening levels are around 60 dB above the average speech threshold, only the upper case of (2.57) is used [114].

SSNR and BSD are used because they give high correlation with subjective distortion measures for speech coders at 0.77 and 0.9 respectively on a normalized scale [114].

2.10.2 Subjective Measures

The purpose of an objective measure is to approximate a subjective measure in a fast and repeatable way. In situations where a more accurate estimation of subjective quality is required, or when objective measures are known to be inaccurate, it is necessary to test human subjects under controlled conditions. The ITU-T P.800 [117] standard defines methodology for conducting subjective tests and the conditions required to minimize bias. In this work, the Mean Opinion Score (MOS) is used as a measure of subjective quality.

A subject is provided headphones in a listening room environment and is asked to listen to a set of training sound clips to familiarise themselves with the test. Each training clip is processed in some way for which approximate ratings are provided. The subject might be asked to give their impression of the perceived quality or intelligibility on a five-point Absolute Category Rating (ACR) scale or the level of annoyance on a five-point Degradation Category Rating (DCR) scale, depending upon the objective of

the test. Once trained, the subject listens to a set of clips and rates them accordingly. Each clip is normalized to a constant perceived amplitude defined in ITU-T P.56 [118] to ensure that subjects do not rate according to signal amplitude.

In many cases it is impossible to play an entire corpus of speech signals to a single listener. Listener fatigue [117] can affect results after listening for long periods, although the phenomena has not been quantified. The set of sound clips should therefore be chosen so that equal distributions of processing and speech signal are presented. The duration of subjective tests in this work is usually 20–30 minutes. Subjects may vary in their opinion despite being given training samples for which two calibration methods can be applied. Firstly, unprocessed speech signals, inserted at random into the test set, should receive the highest ratings; if they are consistently lower then all the ratings can be adjusted accordingly. Secondly, a small set of speech signals can be rated by expert listeners, against which the test subjects can be calibrated. A mapping from subject score to calibrated score can then be derived. The collated scores may be represented by a mean opinion score for each category, with additional statistics to analyse the variance of scores.

2.11 Training and Test Corpora

The subset of the APLAWD database [119] used in this thesis contains ten repetitions of five short sentences, spoken by five male and five female speakers, recorded at $f_s = 20$ kHz and 14 bits per sample. Speech and EGG recordings are recorded contemporaneously but without time alignment. The subset of the SAM [120] database, recorded with similar apparatus, contains a long paragraph (~ 150 seconds) spoken by two male and two female speakers. SAM recordings are considered to contain more natural speech and present a more challenging task for a GCI/GOI detector. A subset of the NTT Database [121] contains 3 male and 3 female talkers, each speaking 5 pairs of phonetically-balanced sentences, sampled at $f_s = 16$ kHz and 16 bits per sample. It is used in Section 6.4 only. Care is taken to ensure that each sentence is unique to either training or test data for algorithms that require training.

2.12 Chapter Summary

This chapter has provided a review of the techniques applied throughout this work. The source-filter model is a representation of human speech where an excitation signal, provided by the pseudoperiodic oscillation of the glottis, is filtered by an all-pole vocal tract filter and radiated to a microphone via the lips. This pseudoperiodicity is delimited by the glottal closure instants and is the foundation of glottal-synchronous speech processing, enabling practical applications including dereverberation, time-scale modification, artificial bandwidth extension and data-driven modelling of the glottal excitation. Many existing speech processing techniques cannot fully exploit this pseudoperiodicity. The estimation and inversion of the vocal tract filter to estimate the glottal excitation waveform from recorded speech signals was also reviewed. A number of models exist to model glottal excitation waveforms but many fail by being overly complex or unable to reproduce a signal of sufficient quality. Objective and subjective measures for measuring the quality of processed speech were discussed.

The remainder of this thesis builds upon the techniques presented in this chapter. The detection of GCIs and GOIs from EGG signals, reviewed in Sections 2.7.1 and 2.8.2, is discussed in Chapter 3. The detection of GCIs and GOIs from both clean and reverberant speech, reviewed in Sections 2.8.1 and 2.9.1, is discussed in Chapter 4. Data-driven voice source modelling, presented in Chapter 5, seeks to improve upon the performance of existing glottal models reviewed in Section 2.5. Real-world applications to existing problems in speech processing, reviewed in Section 2.9 are presented in Chapter 6.

Chapter 3

Glottal Activity Detection from Electroglottograph Signals

3.1 Introduction

DETECTION of glottal activity, namely Glottal Closure Instants (GCIs) and Glottal Opening Instants (GOIs), is central to glottal-synchronous speech processing as they delimit pseudoperiodicity in voiced speech. The broad application of glottal-synchronous processing has given rise to a corresponding demand for automatic and reliable detection of glottal activity for which numerous GCI detection techniques have been proposed. The Electroglottograph (EGG) (or Laryngograph) signal [7] is a measurement of the electrical conductance of the glottis captured contemporaneously with speech recordings. The measured EGG signal is proportional to the glottal contact area, whose derivative (DEGG) during voiced speech contains short, high-amplitude impulsive temporal features (spikes) due to glottal closure and smaller features of opposite sign due to glottal opening. The detection of GCIs and GOIs from the EGG signal is considered to be more straightforward than from the speech signal as discussed in Section 2.8.2. It is used extensively as a reference for speech-based GCI detection algorithms [77, 122–125]. Further uses are found in the analysis of pathological speech, including types of dysphonia [3], vocal fold impact stress [126] and essential tremor [127].

Many approaches analyse the EGG by searching for impulsive features in DEGG [68, 128–130] and compare their amplitudes with thresholds to obtain an estimate of glottal activity during voiced speech. Recent approaches have applied multiscale analysis to detect glottal activity as singularities in the EGG signal [131] and speech signal [132]. Existing techniques are, however, often prone to errors around the end of voicing. The Singularity in EGG by Multiscale Analysis (SIGMA) algorithm, presented in this chapter, benefits from the use of multiscale processing but it extends the approach by performing spike detection on the multiscale product using a group delay method [78] which circumvents the need for thresholding. The robustness of the approach to false detections is further enhanced by Gaussian Mixture Modelling (GMM) [133] which is used to remove detections with unlikely features. The proposed method provides accurate GCI and GOI detection. Additionally, the algorithm makes no assumptions about the nature of the EGG signal other than the bounds on the range of glottal frequency and open quotients [134]; SIGMA may therefore have many further uses as it is also suitable for singularity detection in applications outside the field of speech processing.

This chapter is organized as follows: Section 3.2 formulates the problem of EGG-based GCI/GOI detection and the methodology employed by some existing algorithms. Section 3.3 describes the SIGMA algorithm for the detection of GCIs/GOIs from EGG signals. The proposed algorithm is compared with existing techniques and evaluated in Section 3.4. A chapter summary given in Section 3.5.

3.2 Problem Discussion

A voiced speech signal, its corresponding time-aligned EGG signal and the EGG derivative are shown in Figure 3.1. Time alignment is achieved by assuming that the lip-microphone propagation distance plus an estimate of the length of the talker’s vocal tract is a constant value, then subtracting the corresponding delay.¹ We define a positive EGG signal to correspond to high glottal contact area, giving positive- and negative-going transients for GCIs and GOIs respectively.

¹A discussion regarding the accuracy of time-alignment is given in Section 4.4.5.

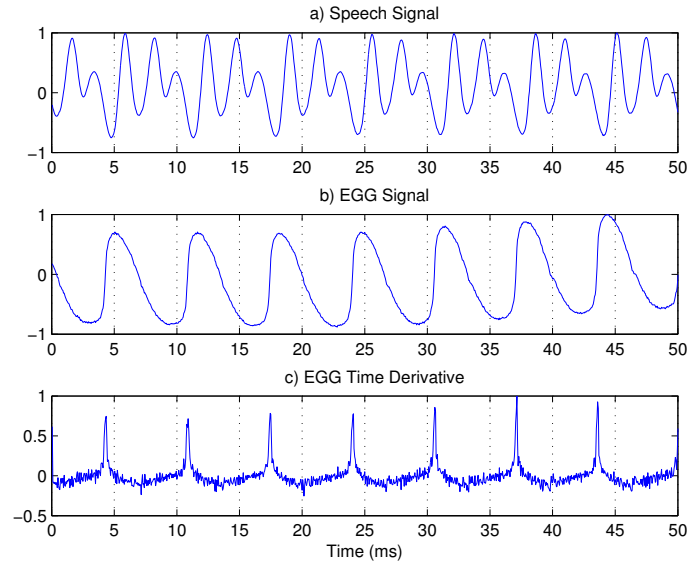


Figure 3.1: Speech signal (a), the corresponding EGG signal (b) and the EGG time derivative (c) for /a/. Negative peaks due to glottal opening are weak in (c).

3.2.1 Defining the GCI

The exact definition of GCIs and GOIs from the EGG signal varies widely in the literature. Early approaches applied a *thresholding level* [135] to the EGG signal, where the GCI is defined as the instant in time when the amplitude equals a percentage of the maximum value. Suggested thresholds are 50% for normal voice and 35% for relaxed voice. Such methods are imprecise when compared with invasive measures of glottal contact area or volume flow [136]. It was observed in [68] that the derivative of the EGG signal, DEGG, displays a strong positive peak at the GCI and a spread negative peak at the GOI. The detection of GCIs by peak detection on the DEGG signal is employed in [128, 130, 137] and it was further noted in [129] that peaks due to GCIs are either precisely located in time or occur as two closely-spaced *redoubled* peaks. More recent approaches calculate the multiscale product of wavelet spectra on the EGG signal so as to estimate its derivative over a set of dyadic scales and locate converging maxima [136, 138]. The advantage of such an approach is that, while detected GCIs are strongly correlated with those from DEGG, by considering multiple scales it exhibits improved robustness to noise.

The definition of the GCI that displays the greatest consistency in the existing

literature is the centre of strong positive-going impulsive features in the DEGG signal or multiscale product. In the case of the ‘redoubled’ GCI, it is defined as the centre of energy of the two peaks.

3.2.2 Defining the GOI

A glottal closure instant is usually followed by a glottal opening instant (GOI), which manifests itself as a weaker and often more spread peak of opposite sign in the DEGG [68], and whose amplitude is largely speaker-dependent. The thresholding level approach in [135] detects GOIs as the instant in time following a GCI when the EGG amplitude equals a percentage of the maximum value; this is deemed equally imprecise as GCI detection with the same method [136]. GOI detection from the DEGG a challenging problem because of the low amplitude, or complete lack, of the opening pulses in the DEGG signal [129]. An approach based upon both EGG and DEGG attempted to circumvent this problem by assuming that the GCI always produces an instantaneous peak in DEGG, then finding the instant in time following the GCI when the EGG amplitude falls below $3/7$ its peak value [128]. The multiscale product of the EGG signal [138] produces stronger peaks at the GOI than DEGG as they estimate the derivative over multiple dyadic scales and are therefore more suited to the detection of spread peaks.

The exact definition of the GCI differs very little in the literature. However, there is no universally accepted definition of the GOI [80]. In this work, the GOI is defined as the centre of negative-going peaks in the DEGG or multiscale product. Differences between the definitions of the GOI when considering both the EGG and speech signals are discussed in Section 4.4.4.

3.2.3 Errors in Existing Approaches

The definition of the GCI and GOI is relatively straightforward for modal voice speech. Many algorithms give examples of the DEGG and multiscale product applied to the EGG signal but many fail to state the exact methodology for estimating time instants. This is particularly problematic during weakly-voiced speech, and in particular at the offset of

voiced speech, where the GCIs and GOIs become more difficult to detect. Such problems are not known to exist at voiced onsets. Qualitative descriptions of the EGG signal's behaviour in these regions has received some attention but has not knowingly been addressed in any algorithms. This subsection gives detail on the problems associated with GCI/GOI detection in these regions.

Two existing approaches whose workings are well-documented include HQT_x (High Quality Time of excitation) and TXGEN (Time of eXcitation GENerator) [130], against which the performance of SIGMA is evaluated in Section 3.4. HQT_x uses two derived functions: DEGG and an estimation of instantaneous gradient. A threshold function varies dynamically with the EGG signal, whose minimum is set by periods of silence assumed to lie during the first and last 20 ms of the EGG recording. The instants of time when the DEGG and instantaneous gradient exceed this threshold are the estimated GCIs. TXGEN is related to the thresholding level approach in [135] and attempts to detect both GCIs and GOIs. After low-pass filtering the EGG signal at 3 kHz, it is differentiated to find DEGG. High and low thresholds are set by the extrema of the DEGG signal from the entire recording multiplied by constant-valued coefficient. If DEGG passes through both thresholds within a set period of time, an estimated GCI is flagged. A GOI is the point in the EGG signal whose amplitude is equal to the amplitude at the preceding GCI.

Detection Errors

Errors in GCI detection can be divided into two categories [77]: *False alarm* errors are made when more than one GCI is detected within a reference cycle; *Miss* errors are made when no GCI is detected within a reference cycle (GOI errors are treated in the same manner). Errors occur when certain types of EGG signal, discussed in the following sections, cause a poor estimate of the signal thresholds described in 3.2.3.

‘False Alarm’ Errors

It has been shown that, for modal voiced speech, the frequency of oscillation of the glottis and the open quotient are dependent on phoneme and voice quality [7, 128]. Studies have

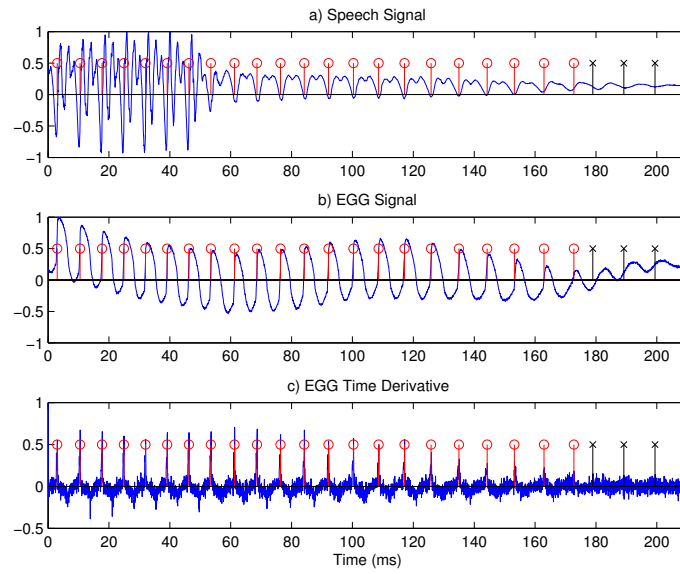


Figure 3.2: A speech signal (a), EGG signal (b), its time derivative (c) and HQTx GCI estimation markers at the end of a voiced speech segment, /u/, exhibiting ‘breathy offset’ (cycles 8-21) and briefly ‘breathy voice’. The first 22 GCIs are identified correctly (marked ‘o’) but the last 3 (marked ‘x’) are erroneous.

further revealed that, for a given talker, the difficulty of detecting glottal closure is largely independent of the sound produced but that interesting effects occur at the boundaries of voiced/unvoiced speech, noting in particular [139]:

1. “Vocal fold vibration does not stop abruptly at the end of voicing, but slowly decays as the vocal folds come to a rest position,” and,
2. “It is possible for vocal fold vibration to continue without the generation of any significant energy,” termed ‘breathy offsets’ [140] .

This is examined in greater detail in [140] where a third phenomenon is observed at the end of voicing:

3. “A persistence of energy in the speech waveform after the EGG waveform has dropped virtually to zero,” termed ‘breathy voice’.

In the case of breathy offsets, GCIs can be detected from the EGG long after the speech amplitude has significantly diminished as the EGG signal remains modal, with increasing open phases that result in a breathier sound [140]. This is demonstrated in Figure 3.2,

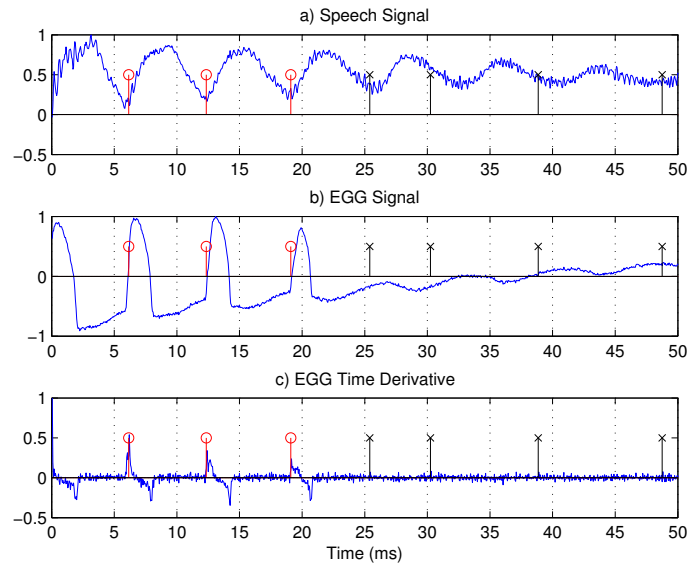


Figure 3.3: A speech signal (a), EGG signal (b), its time derivative (c) and HQT_x GCI estimation markers at the end of a voiced speech segment, /ɪ/, exhibiting ‘breathy voice’. The first 3 GCIs are identified correctly (marked ‘o’) but the last 4 (marked ‘x’) are erroneous. Negative peaks due to glottal opening are significant in (c).

showing 14 cycles of breathy offset terminating in breathy voice when EGG signal finally loses modality and ceases to oscillate with a regular period.

In the case of breathy voice, observed throughout case (3) and at the very end of case (2), the glottis is ‘flapping in the breeze’ [141] with insufficient contact to register on the EGG waveform. As described in [142], “If the glottis does not shut quickly enough...no vocal wave is generated in the supraglottic cavity,” and is demonstrated in Figure 3.3. In both cases a number of erroneous GCIs are detected by HQT_x during segments of breathy voice (x) until its dynamic threshold is no longer exceeded. These errors also often occur at erratic intervals. For the hand-labelled reference, marked ‘o’, the labeller would not mark any GCIs where there is no visible instant defining the periodicity, as would be the case with all instances of breathy voice.

Breathy voice represents a natural transition from modal voiced speech to unvoiced or silence [140]. It is further noted that this usually lasts for just a few cycles of speech but erroneous estimates by a GCI detector during these segments can cause significant problems for glottal-synchronous algorithms. For example, a pitch tracker [143] that cal-

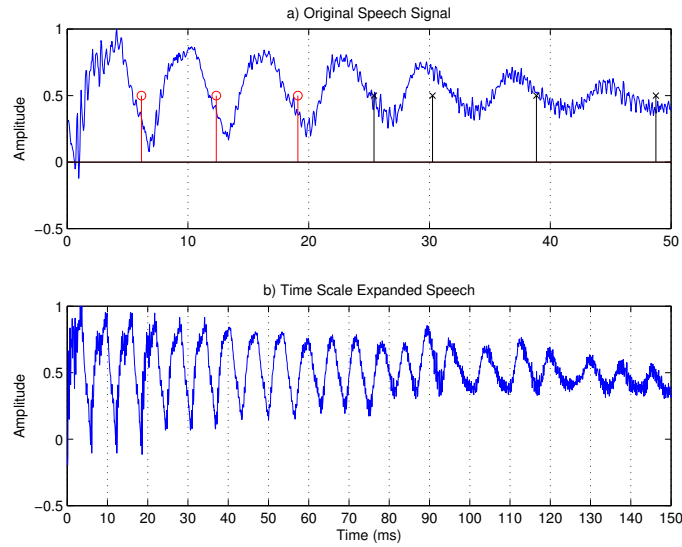


Figure 3.4: a) Original Speech signal with correct GCIs (marked 'o') and false alarm errors (marked 'x') and b) time scale expanded by three times with the PSOLA Algorithm. Voiced cycles are copied and concatenated to increase duration; this works well for modal speech but fails when GCIs are detected in the wrong location.

culates pitch on a cycle-by-cycle basis will give highly erratic results. Glottal-synchronous speech processing algorithms such as prosodic speech modification [144], speech dereverberation [95], speech synthesis [99] and voice source modelling [145] all rely upon the manipulation of individual cycles of speech. Any fricatives or plosives following segments of voiced speech will be treated as periodic, giving rise to particularly annoying artefacts [84].

An example is shown in Figure 3.4 where HQT_x is used to drive the PSOLA algorithm [144] to increase the duration of a speech signal by three times without affecting prosody or formant structure. Applications for increasing the duration of a speech signal include enhancing intelligibility and lip synchronization in motion video. It is achieved by repeating cycles of voiced speech and concatenating them with an estimate of the correct period as shown in the first 70 ms of Figure 3.4 b). Unvoiced speech and voiced-unvoiced transitions do not exhibit such periodicity so a common approach is to leave these segments unmodified [84]. This is not the case due to the erroneous detections at the voiced-unvoiced transition from 70-150 ms, leading to strange artefacts that detract from the otherwise natural sound of the processed voiced speech segments. A detailed investigation into the need for accurate voicing detection when only the speech signal is

available can be found in Section 6.3.

Sudden changes in EGG amplitude can also cause false alarm errors in dynamic threshold-based algorithms if the threshold is too low. A further problem with dynamic thresholds arises when GCIs have slow rise times [131], causing not a spike but a spread pulse in the DEGG. In this case, we define the GCI as the centre of energy of the pulse.

‘Miss’ Errors

A common feature at the end of voiced segments is a reduced EGG signal amplitude compared with normal modal voice. TXGEN’s thresholds are proportional to the extrema of the entire signal and it is generally not prone to the false alarm errors exhibited by HQT_x. It instead gives miss errors where the EGG amplitude is consistently low, particularly at the very beginning and very end of voiced speech segments. For the majority of glottal-synchronous algorithms this does not pose a significant problem. If, however, the amplitude of the EGG signal momentarily drops below the fixed threshold, TXGEN can miss a small number of isolated cycles which can be problematic for certain applications. Data-Driven Voice Source Modelling [145] discussed in Chapter 5, for example, derives feature vectors from individual cycles of voiced speech which are then analysed to determine models of voice source. A missed GCI results in features being derived from multiple cycles of speech, causing misclassification and distorting the processed signal. HQT_x can exhibit miss errors following a sudden decrease in EGG amplitude due to smoothing of the dynamic threshold that is not employed in TXGEN.

The False Alarm / Miss Tradeoff

In general, HQT_x is prone to false alarm errors, particularly at the end of voiced segments. This is verified in section 3.4; it is further shown that miss errors are far less common. In contrast, TXGEN is generally prone to miss errors with relatively few false alarms; this is also verified in section 3.4. HQT_x fails largely because thresholds are estimated over too short a window and TXGEN because thresholds are based upon single global thresholds for the whole speech utterance. The constant of proportionality used to set the threshold

from signal extrema can be varied in TXGEN's function call. The default was empirically chosen to give the best tradeoff between miss and false alarm errors; a marginally lower value can result in increased false alarms and decreased misses. There is therefore a clear tradeoff between false alarms and misses caused by the thresholding approach employed by the majority of existing algorithms. The severity of this type of error is application-specific but, when used as a reference to evaluate speech-based GCI/GOI detectors, neither should be deemed acceptable. SIGMA instead employs a novel method for detecting GCIs and GOIs that does not use thresholding, circumventing the false alarm/miss tradeoff and providing accurate estimates for the entire EGG signal.

3.3 Glottal Activity Detection with the SIGMA Algorithm

Detection of glottal activity from an EGG signal often involves locating positive-going and negative-going peaks in the EGG derivative. An existing approach employed by the HQT_x algorithm is the detection of spikes in the EGG derivative and a longer-term measure of the change in EGG amplitude.

3.3.1 Multiscale Analysis

Let us consider a generalization of the two-gradient approach used in HQT_x. The dyadic wavelet transform [146] involves iteratively decomposing an EGG signal $\eta(n)$ into decimated subbands; a three-level decomposition is shown in Figure 3.5(a), where the down-sampling and filtering operations split the signal into octave-wide subbands.

The filters $g(n)$ and $h(n)$ have high- and low-pass characteristics respectively. It is shown in [138] that, for singularity detection in EGG signals, each filter in the filterbank should be a first-order differentiation operator at increasing levels of smoothing. A wavelet fulfilling this criteria is described as having one vanishing moment and discontinuities in the input signal are seen as converging maxima across scales $d_j(n)$ [147]. A derivative-of-Gaussian (dG) approximation with cubic spline wavelet decomposition filters is used in [148] and [138] which provides the differentiation and smoothing we require. However,

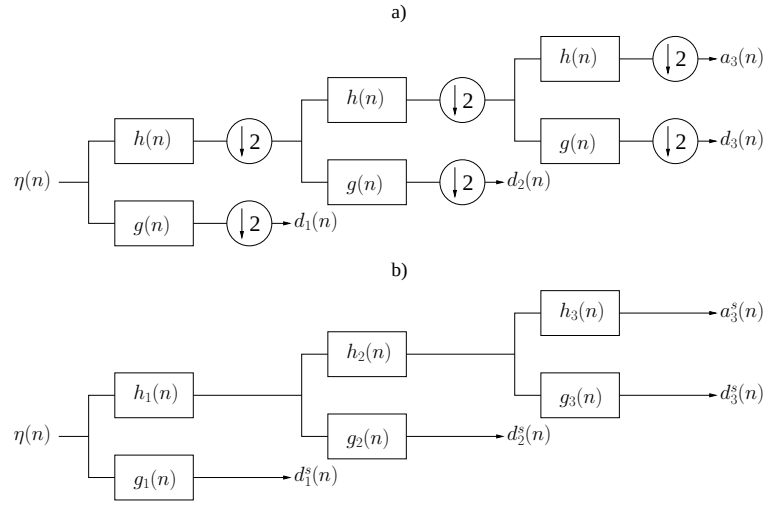


Figure 3.5: Three-level dyadic signal decomposition on a signal $\eta(n)$ into detail, $d_j(n)$, and approximation, $a_j(n)$, signals. a) is the Dyadic Wavelet Transform (DWT), and b) the Stationary Wavelet Transform (SWT), an overcomplete version of the DWT useful in the detection of discontinuities.

an arbitrary number of filters exist which fulfil the same criteria. A number of derivations can be found in [149] but give little idea as to their use in the detection of singularities. In order to determine the relative performance, the proposed algorithm was run with five different sets of decomposition filters. Section 3.4.3 presents a performance comparison between the chosen wavelet, whose filters are shown in Figure 3.6, and the popular cubic spline dG wavelet.

The dyadic wavelet transform is dyadic in both scale and time. Only scale is of interest in singularity detection, so we do not decimate as shown in Figure 3.5(b). Instead, the filters $g(n)$ and $h(n)$ are upsampled by 2 at each iteration to implement the change of scale to form $g_j(n)$ and $h_j(n)$ at scale j . This overcomplete representation of a signal is discussed in detail in [147] and is given many names including: *Stationary Wavelet Transform (SWT)*, *Algorithme à Trous* (Hole Algorithm), *Redundant Wavelet Transform (RWT)* and *Undecimated Wavelet Transform (UWT)*. The signal's length remains unchanged throughout the filterbank tree, allowing simple sample-by-sample multiplication of the signal at different scales to find converging maxima.

Denote the wavelet $\psi_s(t) = (1/s)\psi(t/s)$, where $s = 2^j, j \in \mathbb{Z}$. The SWT of the

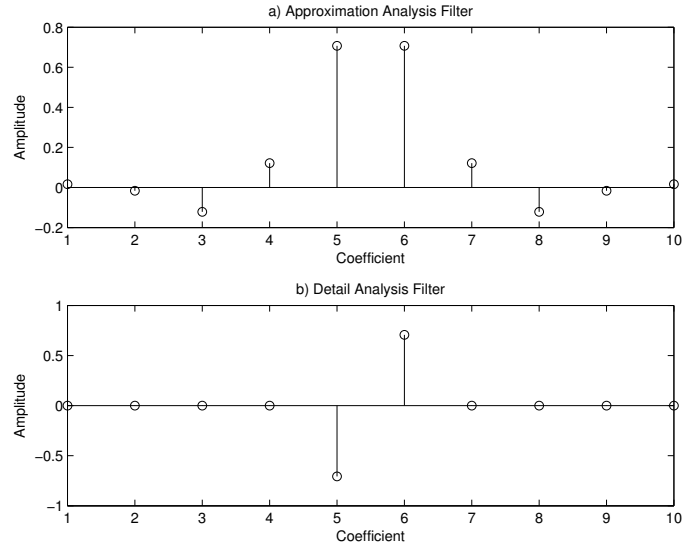


Figure 3.6: Approximation (a) and detail (b) analysis filters for multiscale analysis. Iterating these filters through a dyadic filterbank constructs a biorthogonal spline wavelet with one vanishing moment.

EGG signal at scale j is

$$d_j^s(n) = W_{2^j} \eta(n), j = 1, 2, \dots, J - 1, \quad (3.1)$$

where $J = \log_2 N$, plus the remaining coarse scale information denoted $a_j^s(n)$. This is a straightforward linear filtering operation

$$d_j^s(n) = W_{2^j} \eta(n) = \sum_k g_j(k) a_{j-1}^s(n - k), \quad (3.2)$$

where $d_j^s(n)$ is the SWT of $\eta(n)$ at scale j and a_{j-1}^s are the approximation coefficients at scale $j - 1$. The multiscale product, $p(n)$, is formed by

$$p(n) = - \prod_{j=1}^{j_1} d_j(n) = - \prod_{j=1}^{j_1} W_{2^j} \eta(n) \quad (3.3)$$

where it is assumed that the lowest scale to include is always 1. The sign of $p(n)$ is inverted compared with a DEGG using the chosen wavelet, hence a minus sign is included to maintain the convention. The de-noising effect of $h(n)$ at each scale in conjunction with the multiscale product means that $p(n)$ is near-zero except at discontinuities across

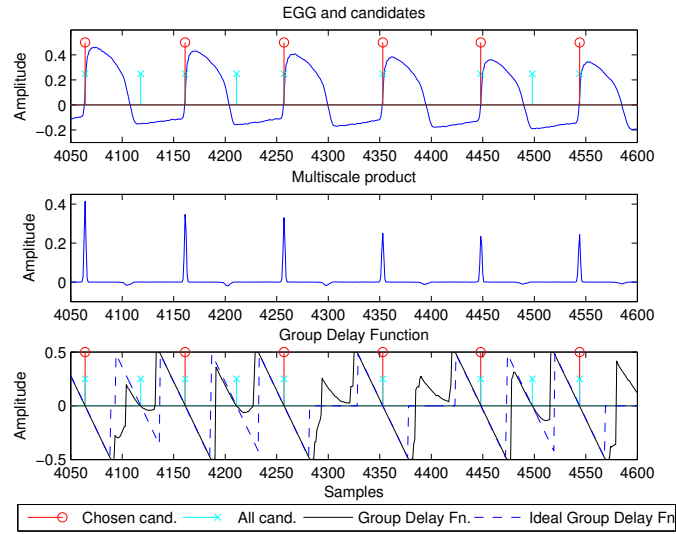


Figure 3.7: EGG waveform, multiscale product and Group Delay Function for GCI detection. Candidates are marked ‘×’ and chosen candidates are marked ‘○’. The ideal slope, marked in a dashed line on the lowest plot, is the slope which would exist if the candidates were perfect impulses.

the first j_1 scales of $\eta(n)$ as depicted in Figure 3.7(b), allowing better identification of discontinuities than the DEGG. The function $p(n)$ can be half-wave rectified to contain peaks pertaining only to GCIs, $p^c(n)$, or GOIs, $p^o(n)$, which aids the group delay function in the following step. The value of j_1 is limited by J , but it is often no greater than $j_1 = 5$ as the region of support (RoS) of $h_j^s(n)$ and $g_j^s(n)$ becomes prohibitively large, demanding high processing resources and smoothing adjacent discontinuities. $j_1 = 3$ is deemed a good compromise [148].

3.3.2 Group Delay Function

A group delay function (GD) [78] can be used for detection of peaks in linear prediction residuals of speech and can be applied to locate spikes in any signal if their minimum separation, T_{min} , is known. Consider the multiscale product, $p^c(n)$, and an L -sample windowed segment beginning at sample n ,

$$q_n^c(l) = w(l)p^c(n + l) \text{ for } l = -L/2, \dots, L/2 - 1. \quad (3.4)$$

The Fourier transform of $q_n^c(l)$ at a frequency $\omega = 2k\pi/L$ is

$$Q_n^c(k) = \sum_{l=0}^{L-1} q_n^c(l) e^{-j\frac{2\pi}{L}lk} \quad (3.5)$$

where k can vary continuously. The group delay of $q_n^c(l)$ is given by [150]:

$$\tau_n^c(k) = \frac{-d \arg(Q_n^c)}{d\omega} = \Re \left(\frac{\tilde{Q}_n^c(k)}{Q_n^c(k)} \right) \quad (3.6)$$

where $Q_n^c(k)$ is the Fourier transform of $q_n^c(l)$ and $\tilde{Q}_n^c(k)$ is the Fourier transform of $lq_n^c(l)$ at frequency $\omega = 2k\pi/L$. If $q_n^c(l) = \delta(l - n_0)$, where $\delta(l)$ is a unit impulse function, it follows from (3.6) that $\tau_n^c(k) \equiv n_0 \forall k$. In the presence of noise, $\tau_n^c(k)$ remains constant but with a degree of additive noise, so an averaging procedure needs to be performed over k ; different approaches are reviewed in [78]. The *Energy-Weighted Group Delay* was deemed the most appropriate [122], defined as

$$\gamma^c(n) = \frac{\sum_{k=0}^{L-1} |Q_n^c(k)|^2 \tau_n^c(k)}{\sum_{k=0}^{L-1} |Q_n^c(k)|^2} - \frac{L-1}{2}. \quad (3.7)$$

Manipulation yields the simplified expression

$$\gamma^c(n) = \frac{\sum_{l=0}^{L-1} l(q_n^c(l))^2}{\sum_{l=0}^{L-1} (q_n^c(l))^2} - \frac{L-1}{2} \quad (3.8)$$

which is an efficient time-domain formulation and can be viewed as the centre of energy of $q_n^c(l)$, bounded in the range $[-(L-1)/2, (L-1)/2]$. The choice of L is related to the minimum f_0 ; sensitivity of the algorithm to L is explored in Section 3.4. The location of the negative-going zero crossings of $\gamma^c(n)$ give an accurate estimation of the location of a spike in a function as depicted in Figure 3.7(c). Additionally, if a spike is spread in time then the group delay method will find its centre of energy, which is particularly useful in the case of the *redoubled* GCI discussed in [131]. The same analysis is applied to $p^o(n)$ to provide $\gamma^o(n)$, whose negative-going zero crossings are GOI candidates.

3.3.3 Candidate Selection

The true GCIs are usually a subset of the negative-going zero crossings of $\gamma^c(n)$, with additional false crossings during unvoiced speech, silence and occasionally between GCIs. Many existing approaches concentrate only on those areas where false candidates are unlikely to occur. The following candidate selection technique aims to remove all false candidates to provide a set of true GCIs throughout an entire segment of speech. Let the number of candidates be $\check{R}^c$ occurring at samples $\check{n}_r^c$, $r = \{0, 1, \dots, \check{R}^c - 1\}$. Three measurements construct a feature vector, $\mathbf{f}_r^c = [f_{r,1}^c \ f_{r,2}^c \ f_{r,3}^c]^T$, from which is derived a feature matrix, $\mathbf{F}^c = [\mathbf{f}_0^c \ \mathbf{f}_1^c \ \dots \ \mathbf{f}_{\check{R}^c-1}^c]$. The features are defined as follows:

- (i) *Consistency of the group delay gradient.* In the case of a Dirac pulse, $\gamma^c(n)$ is a negative unit slope, with a zero crossing at the location of the impulse and width L samples, as shown in Figure 3.7(c). A spread pulse or the presence of noise will cause the slope to deviate from the ideal shape, denoted $I(n)$. The RMS error between ideal and measured is calculated:

$$f_{r,1}^c = \sqrt{\frac{1}{L} \sum_{n=-(L-1)/2}^{(L-1)/2-1} (\gamma^c(n + \check{n}_r^c) - I(n))^2}. \quad (3.9)$$

- (ii) *Peak value of multiscale product's j_1^{th} root inside group delay window.* It is shown in [138] that the j_1^{th} root of $p^c(n)$ provides improved detection of weak discontinuities by reducing the dynamic range between large and small peaks in $p^c(n)$ (in this case $j_1 = 3$). It is suggested in [138] that the j_1^{th} root of $p^c(n)$ can improve detection of discontinuities in $\eta(n)$ as it reduces the dynamic range between impulsive events.

$$f_{r,2}^c = \max \sqrt[j_1]{p^c(n + \check{n}_r^c)}, \quad -\frac{L-1}{2} \leq n \leq \frac{L-1}{2} \quad (3.10)$$

- (iii) *Area beneath multiscale product's j_1^{th} root inside group delay window.* In the case of a spread singularity, the area beneath the multiscale product's j_1^{th} root can provide

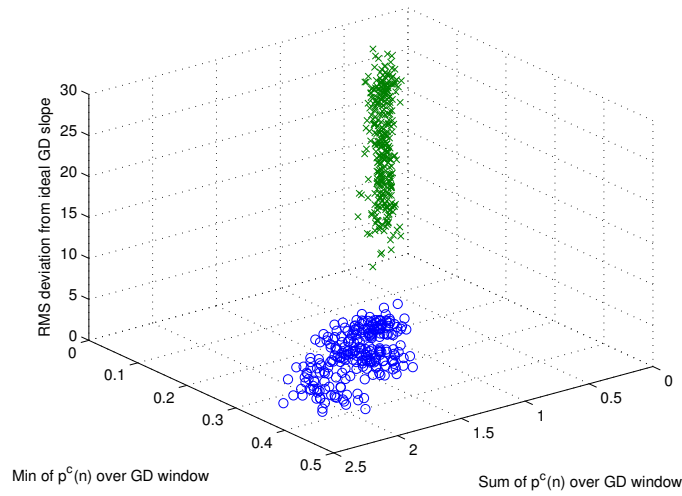


Figure 3.8: Typical distribution of feature vectors for a segment of conversational speech. The chosen cluster, whose members are marked ‘o’ is that whose mean f_3^c is furthest from the origin. Rejected candidates are marked ‘x’.

better discrimination of candidates.

$$f_{r,3}^c = \sum_{n=-(L-1)/2}^{(L-1)/2} \sqrt[n]{p^c(n + \check{n}_r^c)} \quad (3.11)$$

The distributions of the feature vectors are modelled as two multivariate Gaussians using the Expectation Maximization (EM) algorithm [133], initialized with two random data points. Acceptance or rejection is based upon the likelihood of class ω_i , $i = \{1, 2\}$, given feature vector $\mathbf{f}_r^c$,

$$\max_i P\{\omega_i | \mathbf{f}_r^c\}. \quad (3.12)$$

Figure 3.8 shows a typical distribution of the feature vectors for a segment of mixed voiced/unvoiced/silent speech. It has been found empirically that the cluster whose mean f_3 is furthest from the origin is most likely to contain the chosen candidates, marked ‘o’. Rejected candidates are marked ‘x’. The chosen GCI estimates are defined as n_r^c of length R^c . GOIs are calculated in the same way but with reversed signs where appropriate.

3.3.4 Swallowing

The algorithm proposed thus far performs accurate singularity detection on an input signal without considering any characteristics peculiar to EGG waveforms. It is found that in natural conversational speech, singularities are often caused by swallowing and occasionally by electrical interference in the measurement apparatus and are usually single isolated impulse-like signals. Considering a maximum period, T_{max} , all GCIs which are separated from a neighbouring GCI by more than T_{max} are rejected, else they are kept providing $(n_r^c - n_{r-1}^c) < T_{max}f_s > (n_{r+1}^c - n_r^c)$. Experimentation has shown that provided the polarity of the recording is correct, swallowing only causes errors in closure detection so this technique is not applied to opening detection.

3.3.5 GOI Post-Filtering

GOIs n_r^o are detected from $p^o(n)$ using the same approach as applied to GCI detection (with inverted signs where appropriate). However, the energy imparted by glottal opening is often significantly lower than glottal closure, which results in more erroneous GOI candidates. Assuming that a GOI always accompanies a GCI, postprocessing can be applied to use GCI estimates to improve GOIs accordingly. The main cause of error in GOI post-filtering is small perturbations in $p^o(n)$ immediately preceding a glottal closure which triggers a zero crossing in the group delay function. A region surrounding the closure is therefore isolated, limiting the allowed open quotient, Q^o , to the bounds Q_{min}^o and Q_{max}^o . The first candidate which lies within these limits is accepted; if no candidate is found, then one is inserted following the current GCI at the previous open quotient.

The SIGMA system diagram is shown in Figure 3.9. Symmetry can be seen between closure and opening detection up until the postprocessing stage; prior to this point the algorithm need only know the maximum frequency of the singularities to detect and so is suitable for general singularity detection.

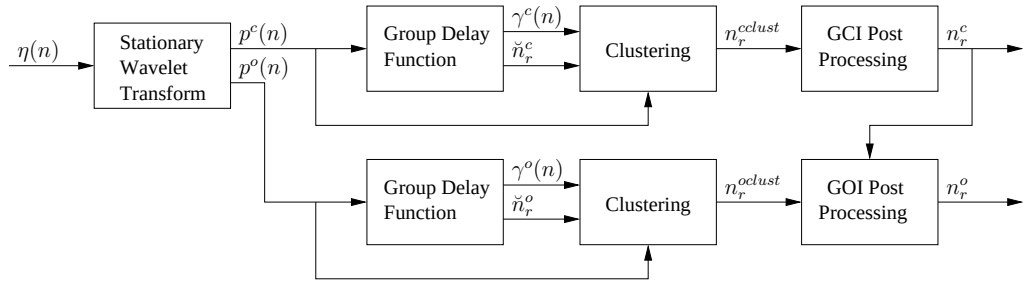


Figure 3.9: SIGMA system diagram. The EGG signal, $\eta(n)$, is decomposed into multiple scales from which the half-wave rectified multiscale product, $p^c(n)$ is derived. Spike detection is performed on $p^c(n)$ by the negative-going zero crossings of the group delay function, $\gamma^c(n)$, at samples $\check{n}_r^c$. Feature vectors derived from the ideal group delay slope and $p^c(n)$ are clustered by an unsupervised EM algorithm to obtain the GCI estimates, n_r^c . Similarly, GOIs are detected using the negative half-wave of the multiscale product, $p^o(n)$. Post-processing is applied to the GCI estimates to remove isolated clicks from sources other than glottal closure to give n_r^c . GOI post processing removes candidates which do not lie within the range of permitted open quotients, using the GCIs as references giving n_r^o .

3.4 Results and Discussion

The SIGMA algorithm has three parameters and these were set as follows:

- T_{min} : the group delay evaluation window size and therefore the maximum frequency of singularities which can be detected. In the case of voiced speech, the maximum glottal frequency is ~ 400 Hz [9] giving $T_{min} = 2.5$ ms.
- T_{max} : the maximum glottal period, so that isolated GCI candidates separated from neighbouring candidates by more than this value are removed in the GCI post-filtering step. A minimum glottal frequency of 50 Hz [9] leads to $T_{max} = 20$ ms.
- $[Q_{min}^o, Q_{max}^o]$: the minimum and maximum open quotients for GOI post-filtering. Their purpose is to isolate a region around a GCI inside which a GOI cannot be detected. They are set at 10% and 90% respectively.

The MATLAB implementation of the chosen biorthogonal spline decomposition filters is called *bior1.5*.

3.4.1 Experiment 1: Evaluation with APLAWD and SAM

The APLAWD database [119] contains speech and contemporaneous EGG recordings of five short sentences, repeated ten times by five male and five female talkers. GCIs and GOIs were hand-labelled on the first repetition of every sentence independently of the algorithms under test according to the criteria defined in Section 3.2.3. Reference GCIs and GOIs are denoted n_r^{cref} , $r = \{0, 1, \dots, R^{cref} - 1\}$, and n_r^{oref} , $r = \{0, 1, \dots, R^{oref} - 1\}$, respectively. A subset of the SAM database [120] contains readings of duration approximately 150 seconds by two male and two female speakers and these were labelled in the same manner. SAM recordings are considered to contain more natural speech with a greater number of swallows and present a more challenging task for a glottal activity detector. The EGG recordings were run through the HQTx (GCI only), TXGEN and SIGMA algorithms and were evaluated by finding the number of estimates per reference cycle then classified as follows, depicted in Figure 3.10:

1. Hit. One estimate per true glottal cycle.
2. Miss. No estimates per true glottal cycle.
3. False Alarm (FA). More than one estimate per glottal cycle.
4. False Alarm Total (FAT). Total number of false alarms (the number of estimates that are not hits).

The measures are defined as:

1. $Hit (\%) = n. hits / (R^{ref} - 1) \times 100$
2. $Miss (\%) = n. miss / (R^{ref} - 1) \times 100$
3. $FA (\%) = n. FA / (R^{ref} - 1) \times 100$
4. $FAT (\%) = n. FAT / R^c \times 100$
5. $Overall (\%) = n. hits / (R^{ref} - 1 + n. FAT) \times 100$

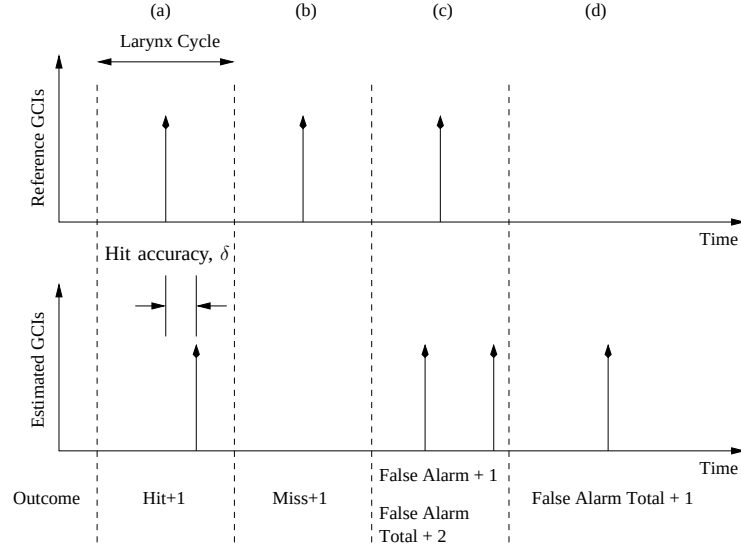


Figure 3.10: Testing strategy. A *hit* (a) is one estimate occurring during a reference cycle. A *miss* (b) is the absence of an estimate per reference cycle. If more than one estimate occurs per reference cycle (c), *false alarm (FA)* is incremented by 1 and total number false alarms in the cycle are added to *false alarm total (FAT)*. If an estimate occurs outside a reference cycle (d), FAT is incremented. Accuracy and bias are the RMS and mean errors between hits and the corresponding reference respectively.

A cycle is bound by $(n_r^{ref} - n_{r-1}^{ref})$ for GCIs, with corresponding evaluation bounds $[\frac{1}{2}(n_{r-1}^{ref} + n_r^{ref}), \frac{1}{2}(n_r^{ref} + n_{r+1}^{ref})]$. GOI cycles are defined in a similar manner. Hit accuracy, σ , and hit bias, μ , are the standard deviation and mean error between all hits and the corresponding ground-truth respectively. The testing strategy is similar to that employed in [77] with the addition of the FAT measure, which counts the *total* number of false alarms as a proportion of total estimates and not the number of reference cycles containing more than one estimate, as a proportion of true glottal cycles. The overall measure provides a single-valued measure of performance by expressing the hit rate as a proportion of all reference cycles summed with the number of non-hit estimates (the FAT).

The GCI results in Tables 3.1 and 3.3 show that SIGMA performs significantly better than HQTx and TXGEN when applied to either database. Notably HQTx is prone to false alarm errors whereas TXGEN is prone to miss errors; this agrees with the qualitative analysis of HQTx's performance in Section 3.2 which showed that it is prone to false alarms at the end of segments of voiced speech. HQTx and TXGEN exhibit much greater

Table 3.1: Closure Performance on the APLAWD Database by HQT_x, TXGEN, SIGMA (dG and bior1.5 Wavelet) Algorithms.

	Hit (%)	Miss (%)	FA (%)	FAT (%)	Hit Acc., σ (ms)	Hit Bias, μ (ms)	Overall (%)
HQT _x	96.47	1.09	2.44	4.73	0.75	-0.01	91.95
TXGEN	94.78	3.47	1.76	2.63	0.45	-0.13	93.27
SIGMA-dG	99.49	0.25	0.26	0.38	0.04	0.01	99.12
SIGMA	99.59	0.26	0.14	0.18	0.04	0.02	99.41

Table 3.2: Opening Performance on the APLAWD Database by TXGEN, SIGMA (dG and bior1.5 Wavelet) Algorithms.

	Hit (%)	Miss (%)	FA (%)	FAT (%)	Hit Acc., σ (ms)	Hit Bias, μ (ms)	Overall (%)
TXGEN	94.63	3.55	1.82	2.05	0.86	-0.05	93.05
SIGMA-dG	99.38	0.30	0.32	0.43	0.24	0.04	98.96
SIGMA	99.47	0.32	0.21	0.24	0.18	0.04	99.23

FAT than FA which suggests that each false alarm is usually followed by successive false alarms within a single reference cycle. SIGMA's miss, FAT and FA measures are broadly similar which tells us that successive false alarms do not usually occur within a given reference cycle and that misses and false alarms have similar likelihood. SIGMA's overall figures of merit are more than an order of magnitude greater than the other algorithms under test.

SIGMA's GCI hit accuracy is in the order of a few samples which agrees with the statement in Section 3.3.3 that the true GCIs are usually a subset of the SIGMA candidate GCIs before clustering. SIGMA and HQT_x hit bias are universally low but TXGEN's estimates tend to occur slightly early. SIGMA's GOI results in tables 3.2 and 3.4 are also encouraging. The reliance upon the estimated GCIs results in similar hit, miss and false alarm rates, with diminished hit accuracy due to the greater difficulty of precisely locating openings. The gap in the overall figure of merit between SIGMA and TXGEN is again more than an order of magnitude.

Table 3.3: Closure Performance on the SAM Database by HQT_x, TXGEN, SIGMA (dG and bior1.5 Wavelet) Algorithms.

	Hit (%)	Miss (%)	FA (%)	FAT (%)	Hit Acc., σ (ms)	Hit Bias, μ (ms)	Overall (%)
HQT _x	95.68	0.27	4.05	14.56	0.37	-0.01	81.85
TXGEN	90.22	9.75	0.03	0.03	1.08	-0.21	90.19
SIGMA-dG	99.27	0.34	0.39	0.44	0.15	-0.05	98.83
SIGMA	99.35	0.50	0.14	0.17	0.16	-0.04	99.18

Table 3.4: Opening Performance on the SAM Database by TXGEN, SIGMA (dG wavelet and bior1.5 Wavelet) Algorithms.

	Hit (%)	Miss (%)	FA (%)	FAT (%)	Hit Acc., σ (ms)	Hit Bias, μ (ms)	Overall (%)
TXGEN	90.44	9.44	0.11	0.14	2.06	-0.11	90.31
SIGMA-dG	98.95	0.34	0.71	0.88	0.29	0.02	98.08
SIGMA	99.23	0.38	0.39	0.50	0.25	0.01	98.74

3.4.2 Experiment 2: Variation in Group Delay Window Size

The group delay evaluation window size was set according to the physical constraints of human speech, whose minimum fundamental period is around 2.5 ms. This experiment assesses the algorithm's sensitivity to variation in the group delay window size on the APLAWD database.

The results presented in Figure 3.11 show that 2.5 ms is indeed an optimal choice of window length. The reliance on GCIs to estimate GOIs means that intuitively the overall, hit, miss and FAT rates should vary in a similar manner which is confirmed by these results. FAT rates increase with decreasing window sizes due to the fact that more negative zero crossings can occur in the group delay function per unit time. In this case the true candidates remain a subset of all candidates, with a number of additional false ones arising. Providing the clustering algorithm can discriminate against the false candidates, those which are correct should always be detected so false alarm rates should therefore

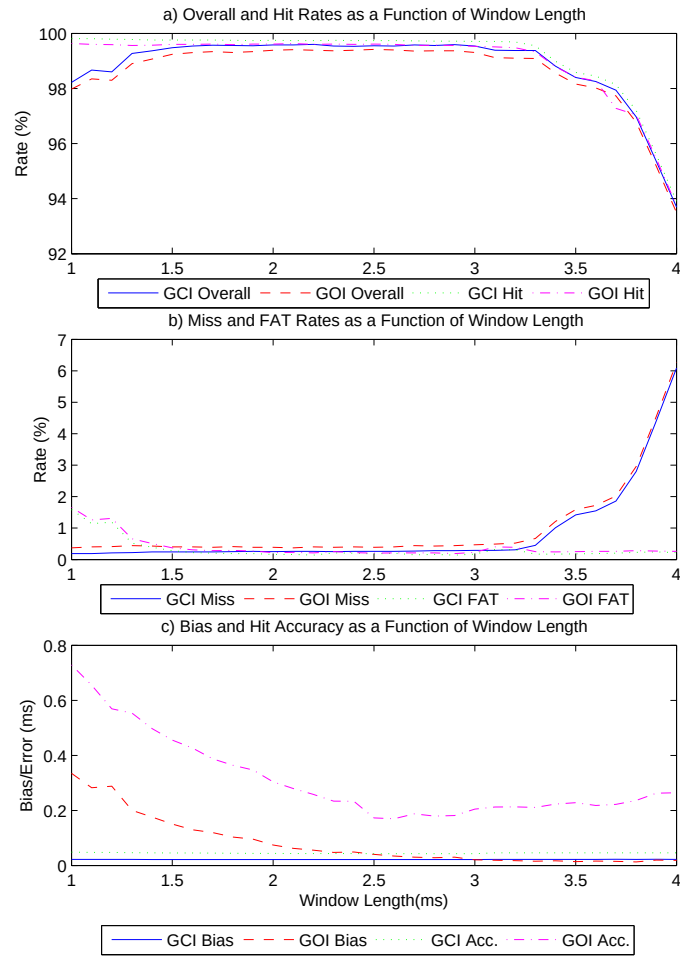


Figure 3.11: The effect of varied group delay window length on: (a) Overall and hit, (b) miss and FAT and (c) bias and hit accuracy. The choice of 2.5 ms from physical reasoning is close to the optimal value.

increase slowly with decreasing window size. Miss rates increase with window size as neighbouring singularities can occur within a single group delay window and reduce the number of negative zero crossings. It becomes impossible for the GMM to find the correct candidates as they are no longer a subset of the candidate set, hence miss rates climb rapidly with increasing window size. GCI bias and hit accuracy are relatively immune to variations in window size, suggesting that providing one candidate occurs per true period, is it statistically the correct choice. GOI bias and hit accuracy are more sensitive, showing the most significant increase with reduced window size. Bias increases monotonically with decreasing window length.

This experiment was repeated for male- and female-only speech. The results, shown in Appendix A, provide similar curves to the previous experiment that employs both genders, the optimum value being shifted up to approximately 3 ms for male voices and down to approximately 2 ms for female. The experiment with mixed male/female speech shows that variation in group delay size does not have a significant effect upon the results in the range of approximately 1.5 to 3.5 ms, hence performance is weakly dependent on gender.

3.4.3 Experiment 3: Comparison with Cubic Spline Wavelet

The derivative-of-gaussian (dG) cubic spline wavelet is the wavelet of choice for multi-scale analysis in [131, 132] and [148]. Experiments with other common wavelets have shown that the biorthogonal spline wavelet, bior1.5, is more effective for EGG analysis with this algorithm. The results in Tables 3.1–3.4 show SIGMA using the dG wavelet (labelled SIGMA-dG) as well as the proposed bior1.5 (labelled SIGMA). The performance of SIGMA is slightly reduced with the dG wavelet, particularly with increased false alarms and increased hit error on the opening tests. Miss rates are slightly reduced but the greater increase in false alarm rate diminishes overall performance.

3.5 Chapter Summary

We have shown that robust detection of GCIs and GOIs from EGG signals is particularly challenging at the transition regions around the ending of voicing. A new method for glottal activity detection from EGG recordings has been presented which is accurate even in these challenging regions. It first detects singularities in the EGG signal by the multi-scale product of three dyadic scales. It then employs a technique based upon the group delay function which detects peaks in the multiscale product. Incorrect estimates are removed by the clustering of three-dimensional feature vectors using the EM algorithm. Post-processing removes isolated GCIs and uses GCIs to aid GOI detection.

A comparison was made between the proposed approach and two popular existing

methods by evaluating their performance against 50 short and four long hand-labelled sentences. An existing testing procedure with some modifications was used, showing GCI and GOI detection rates of 99.59% and 99.47% respectively. The method enables evaluation of speech-based glottal activity detection algorithms and precise estimation of the closed phase for the estimation of glottal volume flow. Furthermore, few assumptions are made about the nature of the input signal. This allows the application of the proposed algorithm to singularity detection in almost any signal provided the minimum separation of singularities is known.

Chapter 4

Glottal Activity Detection from Speech Signals

4.1 Introduction

THE previous chapter discussed detection of GCIs and GOIs from Electroglottogram (EGG) signals. In many practical situations, where such invasive measurements are unavailable, GCIs and GOIs must be detected from the speech signal alone. The focus of this chapter is to explore speech-based GCI/GOI detection and is organised as follows: Section 4.2 discusses existing algorithms in the context of a common framework. The DYNAMIC programming Phase Slope Algorithm (DYPSA) algorithm, which represents state-of-the art GCI detection, is reviewed in Section 4.3. An approach for GCI/GOI detection from clean and noisy speech is presented in Section 4.4, followed by a multichannel extension to DYPSA that is robust to the effects of reverberation. A chapter summary is given in Section 4.6.

4.2 Problem Discussion and Existing Approaches

Speech-based glottal activity detection algorithms can usually be decomposed into one or more of three stages: (a) Preprocessing, to reinforce or identify glottal excitation instants, (b) Event detection, to identify GCIs/GOIs from the output of (a), and (c) Postprocessing, to remove incorrect detections from (b) by applying suitable heuristics. GOI detection is less common than GCI detection as it is a relatively more challenging problem as will be demonstrated in Section 4.4.4.

4.2.1 Preprocessing

The preprocessor normally assumes that the vocal tract transfer function is stationary throughout an analysis window of 20–30 ms. During this time, a widely-used approach is the detection of discontinuities in the linear model of speech production [21] which correspond to the GCIs and GOIs. To this end, linear predictive coding is first applied to a speech signal to estimate the vocal tract filter coefficients. Inverse-filtering then provides a linear prediction residual which contains sparse impulsive events at the instants of glottal closing and opening. An early example of practical applications of LPC in GCI/GOI detection can be found in [54] and has been applied to many more recent algorithms, notably [33, 151] and [77]. Additional model-based preprocessors include homomorphic processing [53], in which the excitation signal is estimated as the signal components that contributes to fast changes in the speech spectrum. Estimation of the vocal tract energy flow provides evidence of the glottal closed phase [152]. Model-based preprocessing is advantageous because it exploits knowledge of the voice to provide a signal that is more straightforward to analyse than the speech signal alone, providing the model is sufficiently well-suited to the speech signal under test.

Preprocessing to identify discontinuities or changes in speech energy include the multiscale product of wavelet coefficients [132], Hilbert Envelope [123], Lines of Maximum Amplitudes (LOMA) [153] and Frobenius Norm [154]. A third type of preprocessor detects periodicity which includes the autocovariance matrix of the speech signal [2], Zero-

Frequency Resonator [155] and Empirical Mode Decomposition (EMD) [156]. These non model-based approaches are advantageous because they are well-rooted in signal processing and are not constrained by any particular speech model.

Preprocessors have received much interest in the literature but a thorough review of the relative approaches has not knowingly been conducted. The DYPSA and YAGA methods discussed in this chapter combine multiple preprocessing approaches as no single method has been shown to be advantageous over all others.

4.2.2 Event Detection

The preprocessor emphasizes excitations by transforming them into either an impulsive event (e.g. LPC residual), a local maxima or minima of a smoothly-varying waveform (e.g. LOMA), or a zero crossing (e.g. Zero-Frequency Resonator). The latter two are relatively straightforward to detect but impulsive events can often be masked by noise or neighbouring events that can render them difficult to detect.

A straightforward technique for peak detection is a fixed threshold based upon a long-term measure of speech amplitude. This is sometimes used for GCI/GOI detection in EGG signals [130] (Chapter 3) but has limited application in speech signals due to the large dynamic range of natural conversational speech. Dynamic thresholds based on short-term averages [151] yield better results but can be shown to sometimes sit on a knife-edge between missing events or detecting false events if the threshold is too high or too low respectively [69].

The group delay method [150] uses a weighted group delay calculated on a sliding window. The negative-going zero crossings of this function have been shown to reliably detect impulsive events in the linear prediction residual [77, 157, 158]. Different weightings are reviewed in [78]. Phase slope projection [77] further improves estimates by detecting missed zero crossings and inserting them at the most likely time instant.

4.2.3 Postprocessing

Event detection is often limited to techniques that do not consider speech models. Postprocessing applies statistical models to a set of ‘candidate’ GCIs/GOIs to remove false detections and keep the true set by applying techniques such as N -best dynamic programming [79, 159] to minimise a cost function or clustering in a feature space [69]. Costs/features can be derived from attributes such as pitch consistency, waveform similarity, energy, multichannel correlation or goodness of fit to voice source models.

A straightforward GOI detector can be achieved with postprocessing by inserting them at a fixed open quotient relative to the detected GCIs [77]. This value can be set as at an empirical mean Closed Quotient (CQ) $\simeq 0.3$ but fails to produce an accurate estimation for natural conversational speech. Objective functions such as minimising residual error [54] or formant modulation [33] can also be used to obtain GOIs.

4.2.4 Failure Modes

Existing GCI/GOI detectors are subject to errors whose principal causes are summarised as follows:

1. Lack of voicing detection, resulting in spurious GCIs/GOIs during unvoiced speech and silence which can be detrimental to some speech processing algorithms.
2. Preprocessing and event detection usually finds large discontinuities in the speech signal that can often be attributed to GCIs, although other events may result in false detections such as those voices where GOIs impart more energy than GCIs. The resulting periodicity will, in this case, be correct although there is a systematic timing error.
3. Postprocessors that rely upon waveform similarity or pitch consistency can be prone to finding a path through the candidates that corresponds to a halving or doubling in frequency. This is a particular problem if the waveform similarity is based upon the speech signal as it contains many oscillations in addition to the wanted fundamental

frequency, f_0 , hence it is prone to giving high waveform similarity for candidates that are not separated by $1/f_0$ in time.

4. Additive noise can cause spurious discontinuities in the speech signal and estimated glottal excitation signal that are not caused by the GCIs and GOIs of the wanted speech signal. Reverberation is particularly detrimental as sound reflected from hard surfaces is difficult to dissociate from the clean speech signal [4, 122].

The bulk of this chapter discusses techniques to address these problems. The Yet Another GCI Algorithm (YAGA) addresses failure modes 1–3 from single-channel measurements and additionally detects GOIs. A multichannel extension to the DYPSA algorithm [77] address the fourth failure mode.

4.3 The DYPSA Algorithm

The DYPSA Algorithm [77] is an example of the preprocessor–event detection–postprocessor architecture and is a basis for the proposed algorithms later in the chapter.

4.3.1 Preprocessor

The preprocessor is a linear predictive analysis of the speech signal and its inversion to find an LP residual from (2.35),

$$E(z) = \frac{P(z)S(z)}{\hat{V}(z)}. \quad (4.1)$$

The LP residual is approximately an impulse train at the GCIs with some additive noise discussed in Section 2.6.1.

4.3.2 Event Detection

The group delay function, $\gamma(n)$, was discussed in section 3.3.2 as technique for locating impulsive features in noisy signals. It was first applied to GCI detection from LP residuals in [157] then investigated in detail in [78] before being applied to DYPSA [77].

Occasionally the group delay function can fail to produce a zero crossing from $e(n)$ at the GCI. The likelihood of such failures can be reduced by identifying those areas where a local maximum is followed by a local minimum without crossing zero. The *phase slope projection* technique then identifies the midpoint of these two points and projects it onto the time axis with unit slope. This time instant is then added to the candidate set, denoted $\check{n}_r^c$, $r = [0, 1, \dots, \check{R}^c - 1]$, from which GCIs n_r^c , $r = [0, 1, \dots, R^c - 1]$ are estimated.

4.3.3 Postprocessor

The postprocessor employs N -Best Dynamic Programming to find an optimal path through a set of candidates based upon five-dimensional cost vectors. The subset of candidates is selected according to the minimisation problem defined as

$$\min_{\mathcal{R}} \sum_{r=1}^{|\mathcal{R}|} \boldsymbol{\lambda}^T \mathbf{c}_{\mathcal{R}}(r), \quad (4.2)$$

where $\mathcal{R}$ is a subset with GCIs of size $|\mathcal{R}|$ selected from all GCI candidates, $\boldsymbol{\lambda} = [\lambda_A \ \lambda_P \ \lambda_J \ \lambda_F \ \lambda_S]^T = [0.8 \ 0.5 \ 0.4 \ 0.3 \ 0.1]^T$ is a vector of weighting factors and $\mathbf{c}(r) = [c_A(r) \ c_P(r) \ c_J(r) \ c_F(r) \ c_S(r)]^T$ is a vector of cost elements evaluated at the r th GCI of the subset, located at sample n_r . The cost vector elements are defined as follows. All are normlized in the range -0.5 to 0.5 :

Speech Waveform Similarity

A cross-correlation estimator is applied to the speech signal as

$$c_A(r) = -\frac{1}{2} \frac{R_{r-1,r}}{\sqrt{R_{r-1,r-1} R_{r,r}}} \quad (4.3)$$

where R_{r_1,r_2} is the autocovariance of speech centered at lags $\check{n}_{r_1}^c$ and $\check{n}_{r_2}^c$. Neighbouring glottal cycles are likely to yield low cost. A high cost is applied to those candidates that are not well-correlated, for example if they occur part way through a glottal cycle.

Pitch Deviation

Pitch deviation cost is a function of the current and previous two candidates, defined as

$$c_P(r) = 0.5 - \exp\left(-(\chi(\Delta_P - 1))^2\right), \quad (4.4)$$

where Δ_P is pitch deviation defined as

$$\Delta_P = \frac{\min((\check{n}_r^c - \check{n}_{r-1}^c), (\check{n}_{r-1}^c - \check{n}_{r-2}^c))}{\max((\check{n}_r^c - \check{n}_{r-1}^c), (\check{n}_{r-1}^c - \check{n}_{r-2}^c))}. \quad (4.5)$$

The nonlinearity in $c_P(r)$ applies relatively small penalties to small deviations in pitch which are expected to occur in natural conversational speech. The constant χ sets the rate of increase of cost with Δ_P which was empirically set to $\chi = 3.3$ to obtain zero cost at pitch deviation of 25%.

Projected Candidate Cost

Those candidates arising from phase slope projection are penalized with

$$c_J = \begin{cases} 0.0, & \text{for group delay zero crossings,} \\ 0.5 & \text{for phase slope projection.} \end{cases} \quad (4.6)$$

Phase slope projection generate spurious candidates in addition to those relating to the the true GCIs.

Normalized Energy

The normalised energy is calculated as

$$c_F(r) = 0.5 - \frac{F(\check{n}_r^c)}{\max_k (F(\check{n}_r^c - k))} \quad (4.7)$$

where $F(\check{n}_r^c)$ is the Frobenius Norm of $s(n)$ in the vicinity of GCI candidate r and $0 \leq k < L$. The sliding window of length $L = 0.016f_s$ is chosen to contain at least one GCI

excitation. The Frobenius Norm method [154] is based upon singular value decomposition (SVD) of the speech signal. Let p be prediction order, K the window length and $s(n)$ the speech signal. $s(n)$ can then be written as the following data matrix:

$$\mathbf{S} = \begin{pmatrix} s_{p+1} & s_p & s_{p-1} & \dots & s_1 \\ s_{p+2} & s_{p+1} & s_p & \dots & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+K} & s_{p+K-1} & s_{p+K-2} & \dots & s_K \end{pmatrix}. \quad (4.8)$$

It is assumed that $m \geq p + 1$ and that $\mathbf{S}$ has full column rank. Orthogonal matrices $\mathbf{U} = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_m)$ and $\mathbf{V} = (\bar{v}_1, \bar{v}_2, \dots, \bar{v}_m)$ decompose S into its singular-value components,

$$\mathbf{S} = \sum_{i=1}^{p+1} \sigma_i \bar{u}_i \bar{v}_i^t, \quad (4.9)$$

where $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{V} \mathbf{V}^T = \mathbf{I}_{p+1}$, and $\sigma_1 \geq \sigma_2 \geq \dots \sigma_{p+1} > 0$, where σ_i are the singular values, ^{T} is the transpose operator and $\mathbf{I}_{p+1}$ is the identity matrix of order $p + 1$. Let a matrix C be the arithmetic mean of the singular squared values,

$$C = \frac{1}{p+1} \sum_{i=1}^{p+1} \sigma_i^2. \quad (4.10)$$

The value of C is a measure of the deviation of the speech data from the linear prediction model, which is greatest during times of significant excitation. A computationally-efficient formulation is used in DYPSA, defined as

$$F(n_r) = \sum_{k=-K}^K \min(H, K - |k|) s^2(\check{n}_r^c - k) \quad (4.11)$$

where $H = 0.001f_s$ and $K = 0.002f_s$. The cost c_F is smallest when a candidate falls in the vicinity of a true GCI, where the short-term signal energy is at a maximum. Those candidates that do not correspond to high energy are therefore penalized.

Ideal Phase-Slope Function Deviation

An impulsive event in the absence of noise causes a zero crossing with negative unit gradient. The LPC residual signal $e(n)$ contains events that are not truly impulsive causing a deviation from the ideal gradient [151]. The ideal phase-slope function deviation cost measures the deviation such that zero crossings close to negative unit gradient are favoured. The cost is set to zero for projected candidates. The cost is defined as

$$c_S(r) = 0.5 + \min(0, \max(\bar{\gamma}(\check{n}_r^c), 1/\bar{\gamma}(\check{n}_r^c))) \quad (4.12)$$

where $\bar{\gamma}$ is the mean value of the group delay calculated over a short window centred on candidate r such that

$$\bar{\gamma}(\check{n}_r^c) = \frac{1}{\iota} \left(\gamma\left(\check{n}_r^c + \frac{\iota}{2}\right) - \gamma\left(\check{n}_r^c - \frac{\iota}{2}\right) \right) \quad (4.13)$$

where $\iota = 0.003 * f_s$, although has been found to be relatively insensitive to this parameter.

4.4 The YAGA Algorithm

The following section presents the Yet Another GCI Algorithm (YAGA) algorithm for GCI and GOI detection from speech signals.¹ It builds on existing techniques, making five novel contributions:

- *Voice source estimation* – two different techniques are applied for the accurate estimation of the voice source signal, $u_D(n)$, instead of the more common linear prediction residual, $e(n)$. This is beneficial because the effects of both glottal closure and opening are present in this signal.
- *Multiscale analysis* – Discontinuities in $u_D(n)$ are detected by calculating its multi-scale product [146], which is impulse-like in the vicinity of GCIs and GOIs.
- *Waveform similarity* – Previous techniques such as [77] propose a speech-based wave-

¹Baba Yaga is a witch-like character in Slavic folklore. It is also a stout produced by a craft brewery in Massachusetts, USA.

form similarity measure between candidates. Speech contains formant oscillations in addition to the wanted fundamental frequency, f_0 , hence it is prone to giving high waveform similarity for candidates that are not separated by $1/f_0$ in time. By estimating waveform similarity on the voice source, those candidates not separated by f_0 receive a proportionately higher cost.

- *Glottal opening detection* – Using the estimated GCIs as a reference, a technique for estimating GOIs is applied. From a candidate set, two dynamic programming stages are performed. Firstly, GCIs are detected as low-cost candidates followed by a low-energy closed phase, which are then removed from the candidate set. Secondly, GOIs are detected by finding the best path through CQ values relative to the previously detected GCIs.
- *Voicing detection* – The postprocessor uses the waveform similarity cost as a measure of voicing that suppresses erroneous detections outside voiced regions.

The proposed algorithm is designed using the preprocessor – event detection – postprocessor framework and is described in the following sections.

4.4.1 Preprocessor

The inversion of speech, which results in an LPC residual $e(n)$, has been shown to work well as a preprocessor for GCI detection as glottal closures are manifest as impulsive events. Glottal opening, which generally imparts smaller amounts of energy, is more difficult to locate in $e(n)$ but constitutes a more significant disturbance in the voice source signal, $u_D(n)$, defined in (2.32). In order to locate the subtle effects of the GOI, a good estimate of $u_D(n)$ is required for which two approaches are considered: Alku's Iterative Adaptive Inverse Filtering (IAIF) [51] and autocorrelation LPC with an enhanced preemphasis filter defined in Section 5.5.1. Both achieve improved estimation of the voice source over conventional inverse-filtering by removing the spectral contribution of the voice source in the recorded speech signal before estimating the vocal tract AR parameters.

Discontinuities in $u_D(n)$ constitute both GCI and GOI candidates. The multiscale

product of the Stationary Wavelet Transform (SWT), $p(n)$, reinforces discontinuities in a signal by calculating its derivative at multiple dyadic scales and locating converging maxima [147]. The technique is described in detail in Section 3.3.1 where it is applied to the electroglottogram (EGG) signal for the detection of glottal closure and opening instants. The same biorthogonal spline wavelet with one vanishing moment and associated analysis filters is used in this algorithm. It is suggested in [69, 138] that detection of impulsive events in the multiscale product due to discontinuities in $u_D(n)$ can be improved by taking the j_1^{th} root of $p(n)$ and half-wave rectifying to give $p^-(n)$. The evaluation procedure in Section 4.4.5 was run using both $p(n)$ and $p^-(n)$, with the latter providing improved identification rate and accuracy.

4.4.2 Event Detection

The signal $p^-(n)$ contains sparse impulse-like features at the location of GCIs and GOIs. In order to locate these impulses, the negative-going zero crossings of the group delay function [78], $\tau(n)$, is determined for $p^-(n)$ as described in Section 3.3.2. As GOIs are to be detected in addition to GCIs, the group delay is reduced from the previously-used 3 ms [77] to 2 ms as shown in Figure 4.1, showing multiscale product (solid blue), group delay function (black dash) with overlaid GCIs (green $\triangle$) and GOIs (red ∇). The group delay function produces a zero crossing at both GCIs and GOIs in (a) but only GCIs in (b), hence a 2 ms window is used for this algorithm. In those cases where a zero crossing is missed, identified by a local maxima that follows a local minima of the same sign, phase slope projection [78] is applied to locate and correct missed zero crossings which provides the complete candidate set n_r^{cand} .

4.4.3 Postprocessing

Dynamic Programming

The postprocessor employed in the YAGA algorithm employs the scheme set out in Section 4.3.3, which applies N -best dynamic programming [79] to find a path that minimizes a set of costs in order to detect GCIs. It differs in respect of an additional cost, c_C , which

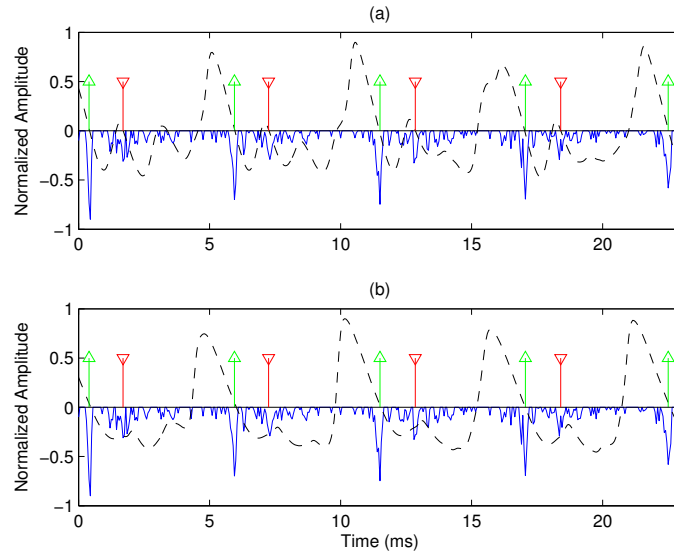


Figure 4.1: Group delay window variation showing multiscale product (solid blue), group delay function (black dash) with overlaid GCIs (green $\triangle$) and GOIs (red ∇) for (a) 2 ms group delay window and (b) 3 ms group delay window.

helps to distinguish between GCIs and GOIs by measuring the energy contained in $u_D(n)$ between successive candidates. Glottal closure causes $c_C(r)$ to be low as the following candidate is likely to correspond to a GOI, so $c_C(r)$ is calculated over the closed phase. The closed phase energy cost is defined as

$$c_C(r) = \frac{\|u_D(n'_r)\|_2}{\max_k \|u_D(n'_{r-k})\|_2}, \quad (4.14)$$

where $\check{n}_r^c \leq n'_r < \check{n}_{r+1}^c$.

Existing algorithms such as [77] calculate a waveform similarity measure on the speech signal to penalize paths between candidates that lie in different locations between successive cycles. Speech signals contain high-frequency vocal tract resonances in addition to the fundamental which can occasionally give erroneously high similarity at incorrect locations. Speech waveform similarity is therefore calculated on the voice source signal which is free from vocal tract resonances, where high similarity only results for candidates that lie in the same location in successive cycles. This reduces the likelihood of erroneous detections and eliminates the problem of detections at multiples of the true fundamental.

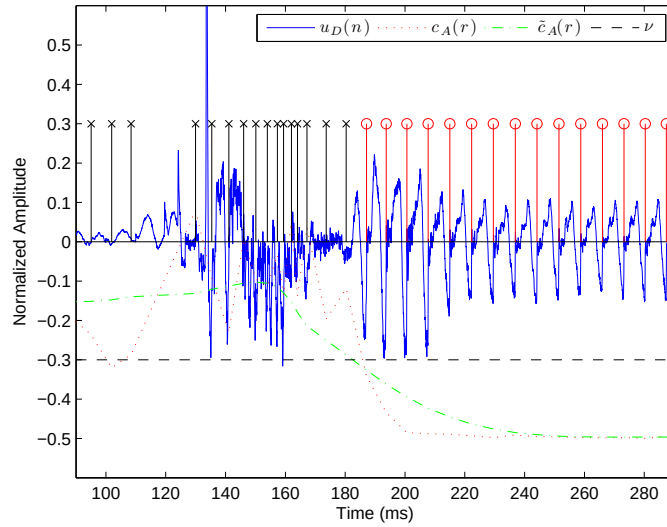


Figure 4.2: Segment of $u_D(n)$ showing silence-unvoiced-voiced transitions, waveform similarity cost, $c_A(r)$, smoothed $c_A(r)$, $\tilde{c}_A(r)$, and threshold ν . $\tilde{c}_A(r)$ provides a good voicing detector; when less than ν , GCIs are kept ($\circ$), else they are rejected ($\times$).

Voicing Detection

The waveform similarity measure is useful not only for eliminating unlikely candidates but it also serves as a reliable measure of voicing. This is required as the algorithm finds GCIs/GOIs during voicing in natural conversational speech but requires an additional voicing detector to remove detections during unvoiced and silent segments.

The duration of voiced segments is relatively long compared with the fundamental period of voicing, T_0 . This permits smoothing of the waveform similarity cost, $c_A(r)$, to help suppress sudden changes which could result in an erroneous voicing decision. Let $\tilde{c}_A(r) = c_A(r) * w(r)$ be a smoothed waveform similarity cost where $w(r)$ is a Hamming window of length 1 ms. A fixed threshold, ν , is used to make a voiced/unvoiced decision,

$$v(r) = \begin{cases} 1 & \text{if } \tilde{c}_A(r) < \nu \\ 0 & \text{otherwise.} \end{cases} \quad (4.15)$$

The parameter ν is set empirically to -0.3. An example of a voiced/unvoiced decision is shown in Figure 4.2, showing $c_A(r)$, $\tilde{c}_A(r)$ and the GCIs that are accepted or rejected based on the value of $\tilde{c}_A(r)$.

GCI Refinement

The zero crossing of the group delay function corresponds to local centres of energy in the voice source signal, which lie in the vicinity of the maximum discontinuity in the voice source. In order to reduce this error, the maxima of the multiscale product lying within 0.5 ms of the zero crossing are identified.

4.4.4 GOI Detection

Definitions of GCIs and GOIs

Glottal closing and opening are not truly instantaneous but phases of finite duration [7], although in general the closing phase is sufficiently short for it to be considered instantaneous. However, there is no universally agreed definition of the precise instants of GCIs or GOIs [80].

There are three main definitions of the opening instant in common use. The first, defined in [54], corresponds to the end of the closed phase when the glottis begins to open, shown by the (o) line in Figure 4.3. There is little evidence of this in the EGG signal of plots (c) and (d) but is present in (b). By delimiting the period of time during which the glottis is closed, the absence of excitation allows closed-phase LPC [54] to estimate more accurately the vocal tract parameters than fixed-frame autocorrelation LPC. The second type of GOI, defined in [7, 139], is the maximum derivative of the EGG signal which corresponds very closely to the end of the opening phase as marked with the (*) line in Figure 4.3. This definition is used extensively to assess open quotients in pathological speech, although it corresponds solely to the maximum rate of change of glottal conductivity and not airflow. The detection of a GOI at the end of the opening phase is not good for applications such as closed-phase LPC as the analysis window should not include glottal excitation. The third type of GOI, defined in the TXGEN algorithm [130], is the point at which the amplitude of the EGG waveform is equal to that at closure following a GCI. Similar EGG amplitude-based thresholds are discussed in Section 3.2. Such a definition is only applicable to EGG signals.

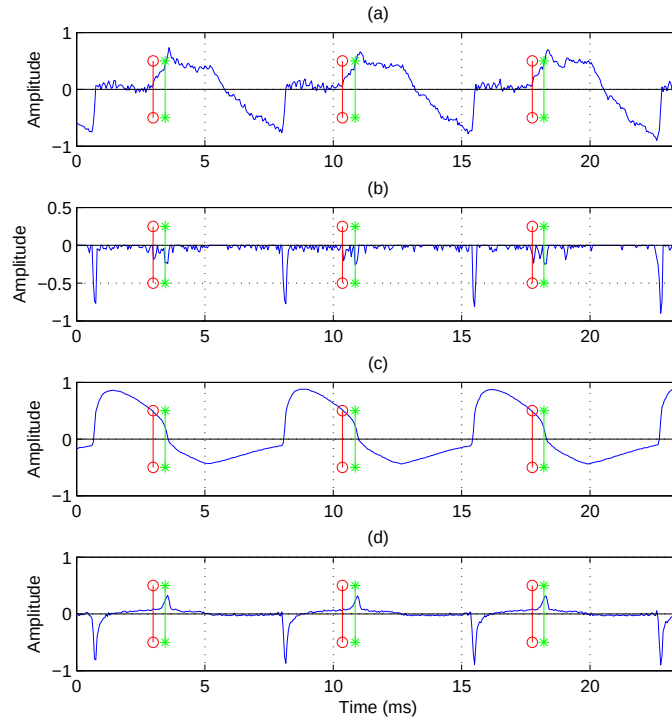


Figure 4.3: Two definitions of glottal opening overlaid on a) voice source, b) multiscale product of voice source, c) EGG and d) multiscale product of EGG. (o) marks the beginning of the opening phase, (*) marks the end of the opening phase.

In this Chapter, the GCIs are defined as large discontinuities in the voice source signal, $u_D(n)$, or equivalently the maxima of the linear prediction residual, $e(n)$. The GOIs are taken as the first definition: the instant of time following the closed phase when the glottis begins to open.

Discriminating Between GCIs and GOIs

The energy imparted into the vocal tract is usually greater at the GCI than the GOI. However, certain breathy voices can display the opposite. Figure 4.4 demonstrates this with the voice source signal for the phoneme /o/ spoken in (a) modal voice and (b) slightly breathy voice. A GCI/GOI detector requires a discriminatory feature that measures closed-phase energy; for a GCI it is low immediately following the candidate and for a GOI it is low immediately preceding. It has been suggested that glottal excitation occurs not only at GCIs

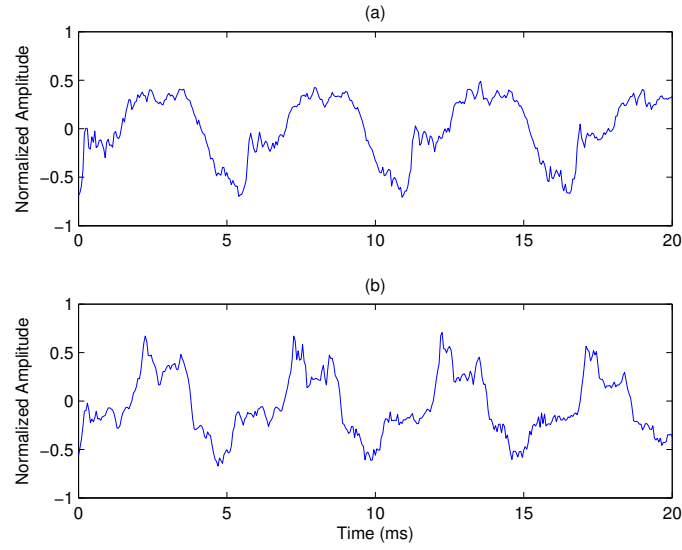


Figure 4.4: Voice source from the phoneme /o/ with a) normal modal voice and b) breathy voice. In b) the energy imparted at the GOIs is greater than at the GCIs.

and GOIs [160]. Recent experiments in [129] demonstrate *double-peaks* where GCI and GOI excitations are manifest not as single events but a pair of closely-spaced events. For the purposes of this work, the centre of energy of double-peaks lying within the support window of the group delay function is taken to be the instant of excitation as in [69].

GOI Detection Strategy

In order to extract GOIs, a new candidate set is first defined by

$$\{\check{n}^o\} = \{\check{n}^c\} \triangle \{n^c\}, \quad (4.16)$$

where $\triangle$ denotes the symmetric difference (union minus intersection) of the two sets. The closed quotient (CQ) relative to n_r^c , Q_r^c , is calculated for each candidate. A dynamic programming algorithm finds the best path by searching for sets of three candidates with CQ within ξ of one another. When multiple sets are found, the one with lowest CQ is selected. A state variable ρ saves the previous good CQ, initialized to 0.2, so that artificial GOIs may be inserted when no suitable candidates are found. Figure 4.5 shows (a) a speech

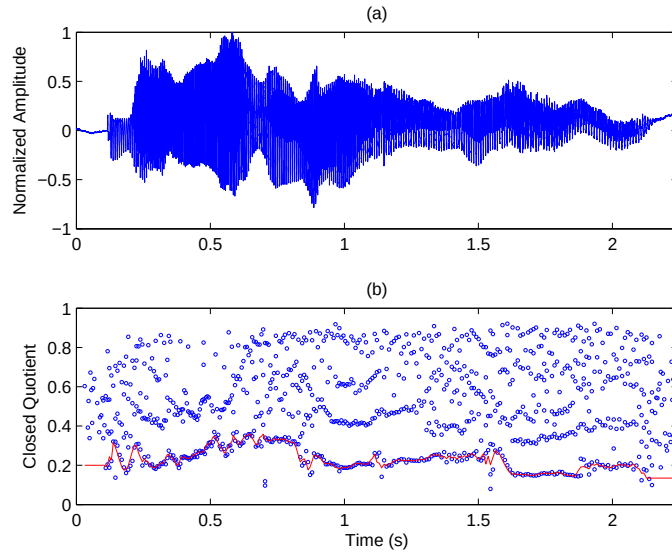


Figure 4.5: (a) Speech signal, (b) CQ of GOI candidates ($\circ$) with best path.

signal and (b) the candidates' CQ ($\circ$) and with the best path overlaid. The estimated GOIs are denoted n_r^o . Visual inspection reveals multiple tracks when excitation is present at both the beginning and ending of the opening phase as discussed in Section 4.4.4. In this chapter, the GOI is marked as the beginning of the opening phase, although by using alternative search criteria different paths may be found.

Figure 4.6 shows a) the glottal volume flow derivative, $u_D(n)$, b) the group delay function, $\tau(n)$, and c) the multiscale product, $p(n)$, with overlaid candidates (cyan) and detected GCIs (green $\triangle$), GOIs (red ∇) following the dynamic programming stage. Candidates corresponding to GCIs show negative-going zero crossings with unit negative slope, whereas GOI candidates would not be identified from $\tau(n)$ without phase slope projection. GCIs are straightforward to identify from $p(n)$ by eye but GOIs are less apparent. The algorithm successfully identifies GOIs as they correspond to a subset of candidates with lowest cost; erroneous candidates with high cost are removed by the dynamic programming.

A system diagram is shown in Figure 4.7.

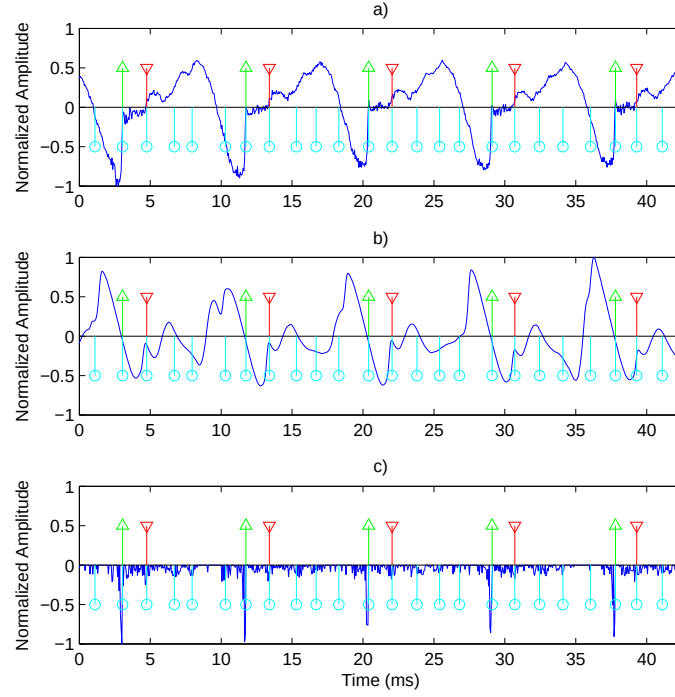


Figure 4.6: a) Excitation signal, $u_D(n)$, b) Group Delay Function, $\tau(n)$, c) Multiscale Product, $p(n)$, with overlaid candidate set (cyan $\circ$) and estimated GCIs (green $\triangle$) and GOIs (red ∇) following the dynamic programming stage.

4.4.5 Evaluation

The algorithm was configured with cost weights $\boldsymbol{\lambda} = [\lambda_A \ \lambda_P \ \lambda_J \ \lambda_F \ \lambda_S \ \lambda_C]^T = [0.8 \ 0.6 \ 0.4 \ 0.3 \ 0.1 \ 0.5]^T$ and CQ tolerance $\xi = 0.1$. The first five elements of $\boldsymbol{\lambda}$ were optimized in [77] and λ_C and ξ were determined empirically. The APLAWD sentence set excluded from training set included 40 sentences and was used to create a hand-labelled reference. Many previous studies use EGG-derived references [77, 132, 155], though the accuracy of EGG-based detectors can vary significantly [69] and many fail to state the algorithm employed. Furthermore, variations in the location of the head relative to the microphone can affect the propagation path of the speech signal and therefore the time delay between the EGG and speech signal. This *bias* can be evaluated between reference and test GCIs and subtracted [155] but it cannot guarantee sample-accuracy. A final disadvantage of an EGG-based reference arises from the discussion in Section 4.4.4, as the GOIs manifest themselves differently in EGG and speech signals. The best ground-truth is therefore deemed to be hand-labelled.

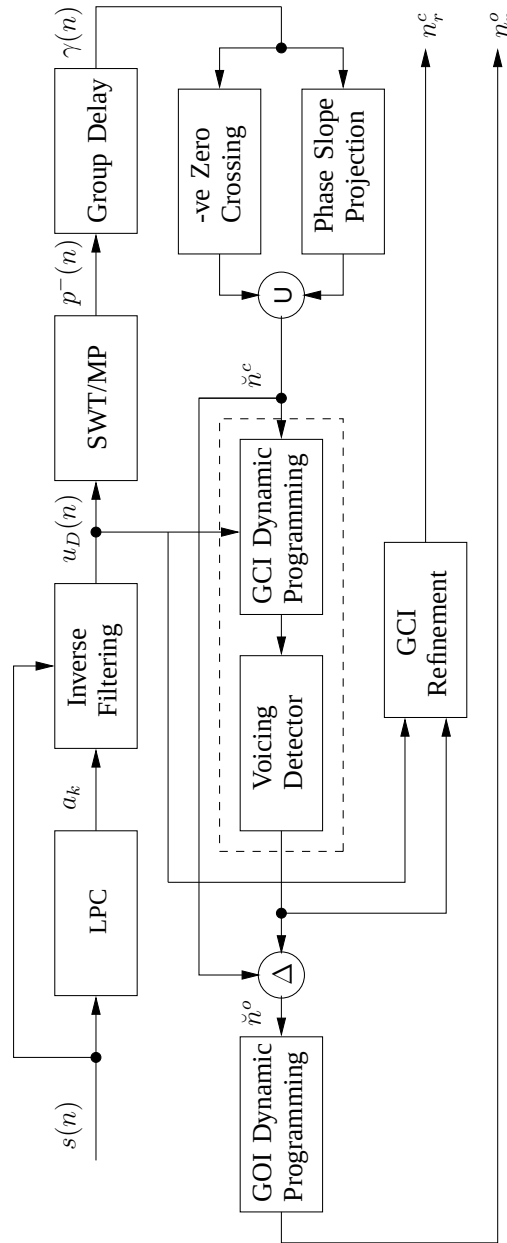


Figure 4.7: System diagram. The preprocessor includes all stages preceding $p^-(n)$, which contains sparse peaks at the locations of both GCIs and GOIs. Event detection is performed by the group delay and associated functions to produce a set of candidate GCIs/GOIs, $\check{n}^c$. A two-stage dynamic programming extracts GCIs, n^c , and GOIs, n^o , from $\check{n}^o$ with optional voicing detection.

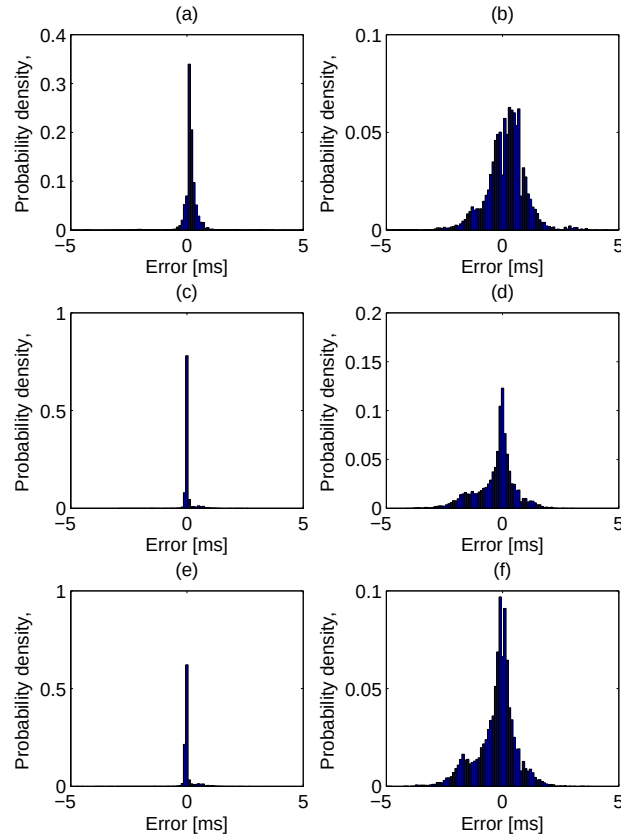


Figure 4.8: Error histograms: (a) DYPGA GCI, (b), DYPGA GOI, (c) YAGA w/ p_{enh} GCI, (d) YAGA w/ p_{enh} GOI, (e) YAGA w/IAIF GCI, (f) YAGA w/IAIF GOI.

A graphical interface was created with which the speech could be labelled by clicking on the signal plots, including simultaneous speech signal, estimated voice source, multiscale product of voice source, EGG, and multiscale product of EGG (calculated as described in Section 4.4.1). Though the EGG signals are not sample-accurate, they nevertheless provide information for the hand-labeller. The GOIs were labelled according to the first definition discussed in Section 4.4.4. In certain cases where the GOI location was unclear, a tool was provided to interpolate the closed quotient from confident GOIs in surrounding cycles. An evaluation strategy identical to that defined in [77] was used with the additional False Alarm Total (FAT) measure defined in Section 3.4, which measures *all* false alarms as a proportion of total candidates, including those between voiced speech segments. This helps to assess the quality of a voicing detector and the ability to suppress multiple false alarms within one reference cycle.

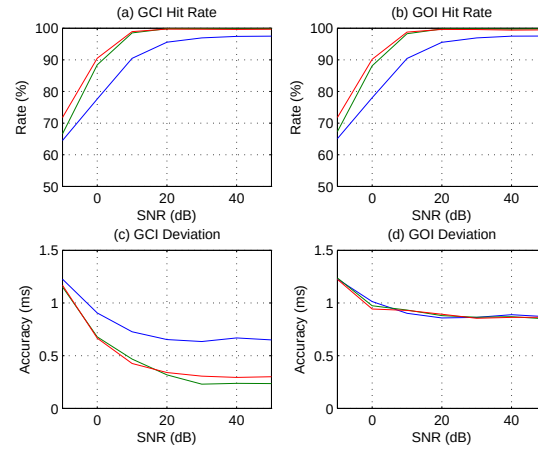
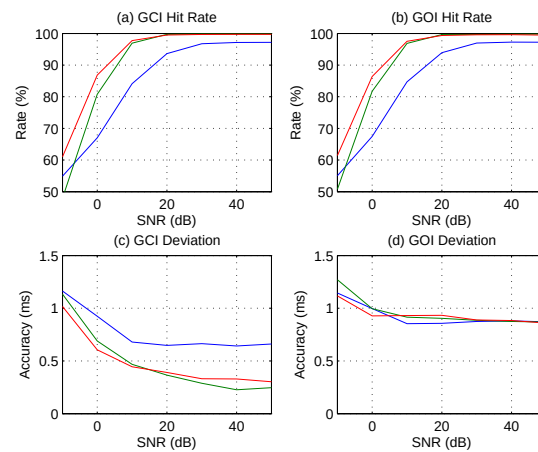
Results on APLAWD Database

The results are shown in Table 4.1, showing significant GCI performance gains over DYPSA with respect to both hit rate and identification accuracy. YAGA with the enhanced preemphasis filter generally has the performance edge, in particular with voicing detection which introduces an excess of misses with the IAIF voice source estimation. However, GOI performance shows little variation between methods. Histograms of identification accuracy shown in Figure 4.8 reveal a slightly different picture. The distribution of DYPSA's GOI error resembles a Gaussian distribution whereas the YAGA variants appear more Laplacian with a sharper central peak. The identification accuracy, as a second-order statistic, cannot describe the dissimilarity between distributions and may unfavourably report the YAGA results. The FAT measure provides insight into the effectiveness of the voicing detector. A marginal reduction in hit rate is juxtaposed with a significant reduction in FAT, making it suitable for those applications which require an additional voicing detector.

The methods under test were evaluated with additive babble and factory noise in Figures 4.9 and 4.10 respectively for SNR values in the range $[-10, 50]$ dB. Speech and noise energy were estimated with ITU-T P.56 [118]. Both YAGA methods generally exceed the performance of DYPSA, displaying little degradation in performance for 10 dB SNR. The introduction of babble noise with a fixed SNR produces lower reduction in hit rate than factory noise in the order of 10%. A possible explanation is that the spectrum of the babble noise mostly lies within the P.56 spectral weighting; to achieve a fixed SNR the factory noise receives a correspondingly higher gain. The reduction in hit accuracy, however, appears to be relatively insensitive to the type of noise. The IAIF voice source estimation exhibits greater noise-robustness, particularly at SNRs of below 0 dB. There is little variation in GOI accuracy between algorithms in the noisy case. There is also little variation in GOI accuracy between 0 and 40 dB, below which SNR GCI and GOI accuracy become broadly similar. The sensitivity of GCI/GOI detection algorithms to additive noise is a relatively sparse subject and, combined with a critical analysis of preprocessors, would make an interesting future study.

Table 4.1: GCI/GOI performance comparison on hand-labelled APLAWD database.

	ID Rate (%)	Miss Rate (%)	FA Rate (%)	FAT Rate (%)	Bias, μ (ms)	ID Acc., σ (ms)
DYPSA GCI	97.67	0.87	1.46	44.41	0.13	0.64
DYPSA GOI	97.65	0.78	1.57	44.38	0.13	0.87
YAGA GCI w/ p_{enh}	99.84	0.06	0.11	45.69	0.02	0.24
YAGA GOI w/ p_{enh}	99.84	0.06	0.11	45.70	0.27	0.84
YAGA GCI w/ p_{enh} + V. det	98.00	1.90	0.07	11.1	0.02	0.24
YAGA GOI w/ p_{enh} + V. det	97.52	2.41	0.07	11.1	-0.23	0.84
YAGA GCI w/IAIF	99.72	0.08	0.20	45.12	0.02	0.30
YAGA GOI w/IAIF	99.51	0.02	0.32	45.14	-0.27	0.86
YAGA GCI w/IAIF + V. det	93.58	6.40	0.03	6.18	0.03	0.25
YAGA GOI w/IAIF + V. det	92.52	7.41	0.08	6.15	-0.27	0.84

Figure 4.9: Hit rate and accuracy performance of methods with additive babble noise with varying SNR. Blue: DYPSA, green: YAGA w/ p_{enh} , red: YAGA w/IAIF.Figure 4.10: Hit rate and accuracy performance of methods with additive factory noise with varying SNR. Blue: DYPSA, green: YAGA w/ p_{enh} , red: YAGA w/IAIF.

4.5 GCI Detection from Reverberant Speech

4.5.1 Problem Formulation

In many modern telecommunication applications, speech signals are obtained in enclosed spaces with the talker situated at a distance from the microphone. The observed speech signal is distorted by reverberation as discussed in Section 2.9.1, diminishing intelligibility [4] and inevitably degrading the performance of GCI identification algorithms. The observation at microphone m is

$$x_m(n) = h_m(n) * s(n), \quad m = 1, 2, \dots, M, \quad (4.17)$$

where $h_m(n)$ is the L -tap impulse response of the acoustic channel between the source to the m th microphone. It has been demonstrated for reverberant speech that reverberation has a significant effect upon the LP residual [161]. Studies on the effect of reverberation on voiced speech LP residuals [93, 94] have further shown that the room impulse response results in additional spurious peaks of similar amplitude to the original excitation peaks, rendering it difficult to distinguish the true GCIs. The impulse-train model of the LP residual, used as a preprocessor to DYPSA, therefore fails in the presence of reverberation. GCI detection from reverberant speech signals forms an important topic of research for the practical applicability of glottal-synchronous algorithms in the future.

Using the characteristics of the linear prediction residuals resulting from clean and reverberant speech discussed in Section 2.9.1 and the properties of DYPSA presented in Section 4.3, the following remarks can be made:

1. The reverberant prediction residual contains many peaks due to the room impulse response, whose amplitudes are comparable to the desired peaks in the clean speech residual. Consequently, the group delay function and the phase-slope projection are likely to produce many erroneous candidates.
2. Peaks of similar amplitude to the true excitation peaks from the clean prediction residual are likely to result in wrong candidates if they both occur in the same group

delay function window.

3. A low-energy voiced speech segment preceded by a high energy component is likely to result in erroneous candidates due to the smearing effect of the room impulse response. Such segments occur, for example, at the end of voiced utterances.

Despite these errors, in multiple time-aligned observations from a beamformer, the peaks due to GCIs are correlated while those due to reverberation are not [4, 162]. This observation has motivated the development of several speech dereverberation algorithms [93, 95, 96] that reduce the effects of reverberation by attenuating such uncorrelated components. This is the motivation for the introduction of multichannel processing within DYPSA to improve its robustness for which two extensions are proposed:

1. A Delay-and-Sum Beamformer (DSB) front-end to the preprocessor. Microphone arrays are known to be advantageous for sound captured in reverberant environments [93] due to the spatial diversity of the Room Transfer Function (RTF) as discussed in Section 2.9.1. A DSB exploits this diversity by time-aligning the recorded speech signals such that cross-channel summation reinforces the wanted signal, which exhibits high interchannel correlation, while attenuating noise and reverberation components that are less correlated.
2. Multichannel DYPSA. The DYPSA preprocessor and event detection can be applied to each of M channels, generating M sets of candidates. Spatial diversity is exploited with this approach by calculating a new cost function based upon the interchannel correlation of candidates and applying it in the postprocessing cost minimization of (4.2).

4.5.2 DYPSA at the output of a beamformer

The output of the DSB can be written

$$\bar{x}(n) = \frac{1}{M} \sum_{m=0}^{M-1} x_m(n - \tau_m), \quad (4.18)$$

where τ_m is a delay to compensate for the time delay of arrival to the different microphones in the array and is assumed to be known. $\bar{x}(n)$ is then presented as a single-channel input to the standard DYPSA algorithm. We refer to this approach as DSB-DYPSA.

4.5.3 Multichannel DYPSA

Multichannel DYPSA (MC-DYPSA) is a novel extension to DYPSA which relies on the correlation of GCI candidates across multiple channels. MC-DYPSA performs preprocessing and candidate generation on each channel independently. The N -best Dynamic Programming (DP) postprocessor selects the most likely GCIs based upon a defined cost function, which is augmented with an additional component that penalizes candidates that are not well correlated across time-aligned channels.

Each channel $m = \{0, 1, \dots, M-1\}$ contains N samples indexed $n = \{0, 1, \dots, N-1\}$ from which a total of $\check{R}_m^c$ GCI candidates are extracted, located at samples $\check{n}_{r,m}^c$, $r = \{0, 1, \dots, R_m-1\}$. Unique GCI candidates (those occurring in at least one channel at the same time) are defined as $\check{n}_r^c = \{\check{n}_{r,0}^c \cup \check{n}_{r,1}^c \cup \dots \cup \check{n}_{r,M-1}^c\}$, so that $\check{n}_r^c$ is the union of the unique GCI candidate sets from all channels. Let $g_m(n)$ be a train of impulses at times corresponding to the locations of GCI candidates for channel m , such that

$$g_m(n) = \begin{cases} 1 & n = \check{n}_{r,m}^c \forall r \\ 0 & \text{otherwise.} \end{cases} \quad (4.19)$$

The mean, $\bar{g}(n)$, of $g_m(n)$ across all channels is a function indicating the number of occurrences of GCI candidates for a given sample n ,

$$\bar{g}(n) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{r=0}^{R_m-1} \delta(n - \check{n}_{r,m}^c) \quad (4.20)$$

where $\delta(n - n_{r,m})$ is a unit impulse function with origin at the candidate r in channel m . Small timing errors can occur in the GCI candidates because of poor channel alignment, phase-slope projection errors and sampling noise (at low sampling frequencies). Therefore a spreading function is applied to $\bar{g}(n)$ so that GCI candidates in close proximity incur a

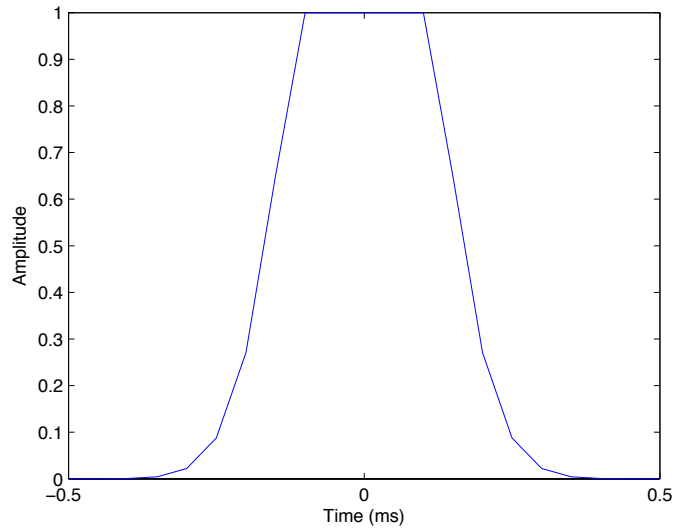


Figure 4.11: Multichannel DYPsA smoothing function – a clipped Gaussian for smoothing the cost component for interchannel correlation.

lower cost than those spread further apart. A clipped Gaussian was found to be a suitable spreading function, as shown in Figure 4.11, denoted by $\Upsilon(n)$,

$$\Upsilon(n) = \begin{cases} ku(n) & 0 \leq |ku(n)| \leq 1 \\ 1 & |ku(n)| > 1. \end{cases} \quad (4.21)$$

where $u(n)$ is a zero mean unit variance Gaussian multiplied by a gain k . It is convolved with $\bar{g}(n)$ to form a new function $d(n)$,

$$d(n) = \bar{g}(n) * \Upsilon(n) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{r=0}^{R-1} \delta(n - \check{n}_{r,m}^c) * \Upsilon(n) \quad (4.22)$$

where $*$ denotes linear convolution. The function $d(n)$ is not bounded in the range $0 < d(n) < 1$ but may exceed 1 depending upon the proximity and height of the samples of $\bar{g}(n)$. Samples for which $d(n)$ exceed 1 are all likely candidates. We next define the inter-channel cost function, $c_I(r)$, such that values of $d(n)$ exceeding 1 are mapped to -0.5 and those in the range $0 < d(n) < 1$ are mapped to $0.5 > d(n) > -0.5$.

$$c_I(r) = \begin{cases} 0.5 - d(n_r) & d(n) < 1 \\ -0.5 & d(n) > 1 \end{cases} \quad (4.23)$$

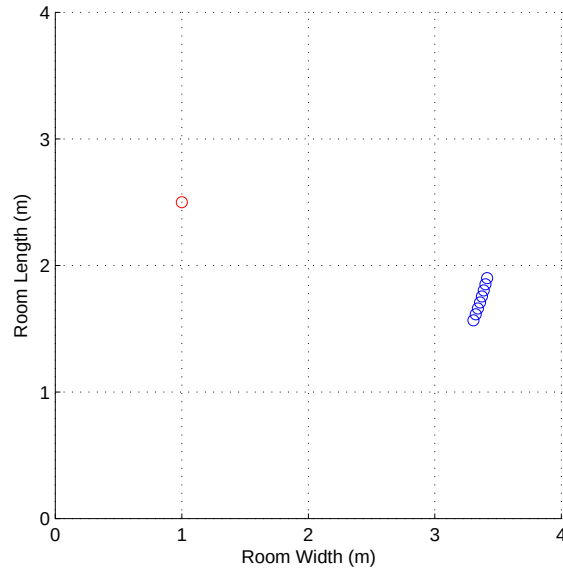


Figure 4.12: Source and Microphone arrangement. The microphone array is 2.5 m from the source on a circular arc to prevent interchannel delay, removing the necessity for time alignment. The array was placed at a slight angle relative to the walls to reduce strong initial reflections.

Note that this cost function is now a function of r and not n for compatibility with the DYPSA DP. This is a linear mapping for $d(n) < 1$, but it is possible a nonlinear mapping may yield better results by penalising low inter-channel correlation and encouraging high inter-channel correlation to a greater degree. The interchannel correlation cost weighting, λ_I , has been empirically set to 0.4.

4.5.4 Evaluation

The value T_{60} is defined as the time for a Room Impulse Response (RIR) to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging $\{100, 150, \dots, 500\}$ ms was simulated using the source-image method [163], containing an array of eight microphones, spaced 50 mm apart, placed on a circular arc 2.5 m from the source so that each channel contained a 2.5 m propagation delay and no interchannel delay (Figure 4.12). Good signal alignment is important and generally requires subsample delays; placing microphones on a circular arc centered at the source alleviates the problem for the purpose of this study. The time-aligned EGG signals were analysed with HQTx [130] to provide reference GCIs.

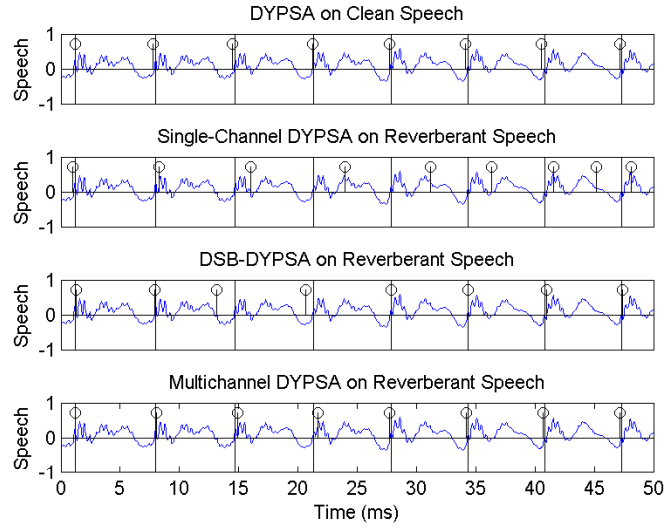


Figure 4.13: Identified GCIs superimposed onto a clean speech signal for a) DYPSA on clean speech. b) DYPSA on reverberant speech. c) DSB-DYPSA on reverberant speech. d) MC-DYPSA on reverberant speech. Reference GCIs obtained with HQTx are represented by solid vertical lines and estimated GCIs are lines ending in a circle.

Experiment 1

A speech file from the APLAWD database was analysed with DYPSA. The sample was then convolved with channel 1 of the microphone array in the $T_{60}=500$ ms case then analysed with DYPSA, DSB-DYPSA and MC-DYPSA. The results depicted in Figure 4.13 show eight reference GCIs derived from the associated EGG signal with HQTx as solid vertical lines and estimated GCIs as short lines terminating in a circle, against the clean speech waveform. DYPSA correctly identifies GCIs with small margins of error when operating on clean speech, but accuracy falls and spurious GCIs increase in the reverberant case. DSB-DYPSA shows improvement with no spurious GCIs but accuracy is significantly lower than the clean case. MC-DYPSA achieves identification on a par with clean DYPSA. This experiment is somewhat idealized which merely demonstrates common errors made by DYPSA and DSB-DYPSA with reverberant speech. MC-DYPSA operating on reverberant speech will not always identify GCIs as well as DYPSA on clean speech.

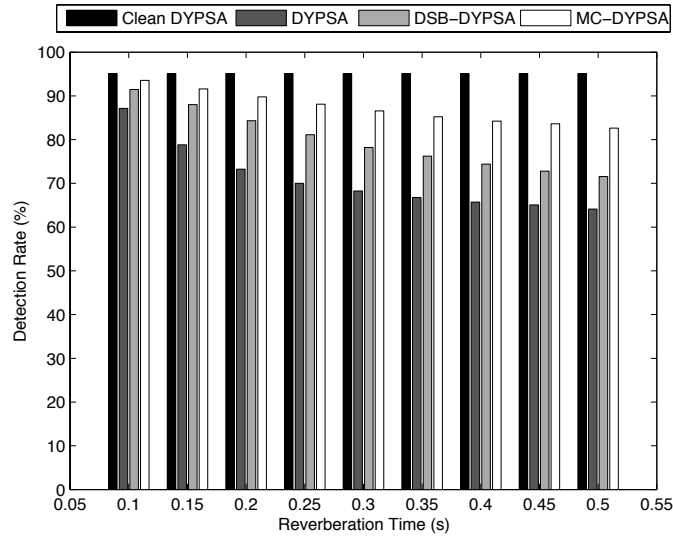


Figure 4.14: Detection rate vs. reverberation time for DYPSA on clean speech, DYPSA on reverberant speech, DSB-DYPSA on reverberant speech and MC-DYPSA on reverberant speech.

Experiment 2

The APLAWD database was convolved with each RIR in turn and analysed with DYPSA, DSB-DYPSA and MC-DYPSA. The results are shown in Figure 4.14, Figure 4.15 and Table 4.2. In all cases, the greatest degradation in detection rate occurs in the lower increments of T_{60} and tails off gently with high reverberation. Single-channel DYPSA shows the worst degradation, dropping by 8% between clean and $T_{60}=100$ ms and 31% at $T_{60}=500$ ms. Multichannel achieves the best with a 12% drop at $T_{60}=500$ ms. Miss and false alarm rates also show significant improvement.

Like detection rate, the greatest degradation in accuracy occurs in the first few increments of T_{60} and tails off with higher reverberation. MC-DYPSA has a higher hit rate so more candidates are included in the calculation of accuracy, causing MC-DYPSA to appear to degrade further than DYPSA and DSB-DYPSA with high T_{60} . Note that hit rate and accuracy from clean DYPSA differ slightly to those given in [77] because the reference GCIs were derived from a newer version of HQTx.

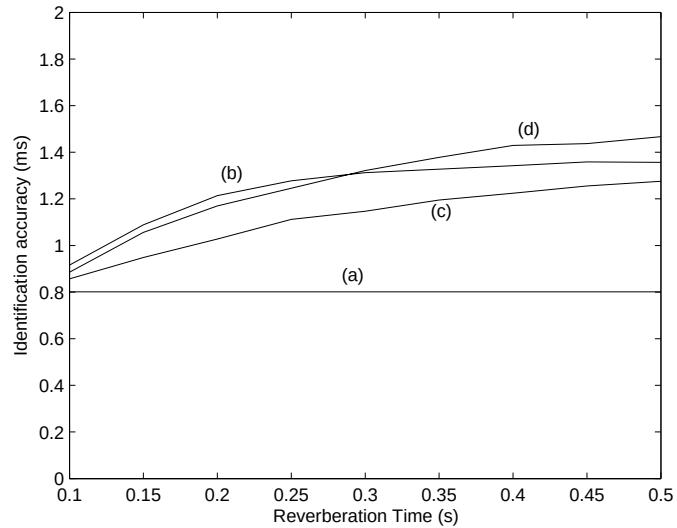


Figure 4.15: Identification accuracy vs. reverberation time for a) DYPsA on clean speech, b) DYPsA on reverberant speech, c) DSB-DYPsA on reverberant speech, d) MC-DYPsA on reverberant speech.

Table 4.2: Performance comparison for DYPsA algorithms on the APLAWD database.

	ID Rate (%)	Miss Rate (%)	FA Rate (%)	ID Acc., σ (ms)
Clean DYPsA	95.1	2.3	2.6	0.80
0.1s DYPsA	87.1	4.1	8.8	0.92
0.1s DSB-DYPsA	91.5	3.3	5.3	0.86
0.1s MC-DYPsA	93.5	2.5	4.0	0.89
0.5s DYPsA	64.1	7.4	28.5	1.36
0.5s DSB-DYPsA	71.5	6.6	21.8	1.27
0.5s MC-DYPsA	82.6	4.1	13.3	1.46

4.6 Chapter Summary

The detection of GCIs and GOIs from recorded speech signals has been reviewed. The DYPSA algorithm represents state-of-the-art GCI detection, in particular with its use of N -best dynamic programming for the removal of erroneous detections. However, evaluation of detection rates against reference GCIs have shown that the approach is not infallible.

The Yet Another GCI Algorithm (YAGA) builds upon the N -best dynamic programming approach and implements enhanced candidate generation using multiscale analysis applied to an estimation of the voice source. Following the detection of GCIs, a novel GOI detector finds the best path through the candidates by searching for consistency in the open quotient. Optional voicing detection suppresses detections during unvoiced speech and silence. Evaluation against hand-labelled data shows a hit rate of 99.84 % on clean speech. The standard deviation of the GCI and GOI error distributions is 0.24 ms and 0.84 ms respectively. Additive noise performance is also promising, with little degradation for all values of SNR > 10 dB.

Environments such as offices often cause significant sound reflection, resulting in reverberation and limiting the applicability of existing GCI detection algorithms in these situations. A microphone array and DSB used as a preprocessor to DYPSA can considerably improve the estimation of GCIs and may provide acceptable results in environments with moderate levels of reverberation. Multichannel DYPSA is a novel extension to DYPSA which uses the correlation of GCI candidates from each microphone in an array to provide highly robust GCI estimation. Though MC-DYPSA contains many parameters that require optimization, preliminary results presented in this chapter suggest that the adopted approach yields very good GCI estimation in highly reverberant environments. With a 0.5 s reverberation time, the identification rate of the proposed algorithm exceeds that of DYPSA by 18%.

Chapter 5

Data-Driven Voice Source Modelling

5.1 Introduction

VOICE source waveform models were reviewed in Section 2.5 and are important in areas of speech analysis, synthesis, coding, recognition and enhancement. Existing models often lend themselves to specific applications and may be partitioned into 3 groups:

Parametric curve fitting models [28,33,34,164] are motivated by observations made by speech scientists from estimated voice source waveforms, for which a combination of polynomial, exponential and trigonometric functions might be used to describe the phases of the glottal cycle. Though conceptually straightforward, the parameters do not necessarily reflect physical or acoustic changes. Parameters are also often assumed independent, allowing them to produce waveforms that do not occur in real observations and rendering their estimation a challenging problem [33,36]. They are also unable to reproduce many of the more subtle components observed in real voice source waveforms. They are nevertheless used extensively in speech synthesis [165] and in areas of recognition [33].

Physical models describe the vocal folds as one or more 2nd-order coupled masses [37–41]. Their ability to self-oscillate in the presence of airflow from the lungs is de-

pendent upon the presence of the vocal tract; this phenomenon of source-tract coupling [12, 166] is considered by only a small number of parametric curve fitting models. However, they incorporate a large number of parameters that are difficult to estimate and require high model orders in order to reproduce subtle components observed in real voice source waveforms. Their use is largely in the synthesis and analysis-by-synthesis of speech [39].

Error minimizing models [23, 42–46] are designed to minimize the error between a recorded speech signal and a resynthesized signal according to an optimization criterion such as mean square error. A combination of noise codebooks, glottal pulse codebooks and long term predictors might be used to reduce error with little or no physical significance. They are used extensively in speech coders where a good representation of all voice signals is required – including mixed and unvoiced speech – although the codebooks themselves may not be optimally designed. Various studies have been undertaken that calculate a mean glottal cycle [53] from a large database of speech, though no known study has yet been undertaken that aims to parameterize the voice source with similar data.

Existing models of the voice source generally impose constraints such as the number of parameters and the effect each parameter has upon the voice source waveform. In this chapter, a new technique termed Data-Driven Voice Source Modelling (DDVSM) is developed that uses machine learning techniques to identify correlation and class separability in a large dataset to extract salient features. This is enabled through the existence of a reliable speech-based GCI detection which segments the voice source waveform into individual cycles; each cycle is normalized in scale and amplitude so as to remove dependence on f_0 and amplitude. Two broad classes of data-driven modelling are considered. Feature Modelling encompasses models that use features used in existing speech processing applications, such as Mel-Frequency Cepstrum Coefficients (MFCCs) [167] and Perceptual Linear Prediction (PLP) coefficients [168] which are used extensively in speech recognition. They are not usually intrinsically invertible although approximations have found uses in areas such as Artificial Bandwidth Extension (ABWE) as discussed in Section 6.4.

Transform Modelling encompasses models employing transforms trained to capture large amounts of signal energy with few independent parameters. Clustering techniques can be applied to both approaches to find clusters existing within the data.

The aim of this chapter is to present DDVSM as a framework for future modelling techniques in addition to discussing specific examples that demonstrate the value of the approach. The remaining sections are organized as follows. DDVSM is formulated as a machine learning problem in Section 5.2, a set of algorithms for generating models are proposed in Sections 5.3 and 5.4, and application examples are given in Section 5.5. The chapter concludes with a summary in Section 5.6.

5.2 Preliminary Processing

The voice source waveform is first conditioned so that machine learning techniques can be straightforwardly applied. Consider a speech utterance, $s(n)$, with which an inverse-filtered voice source waveform, $u_D(n)$, is obtained with Iterative Adaptive Inverse Filtering (IAIF) [51]. Glottal closure instants, n_r^c , are obtained with the YAGA algorithm defined in Section 4.4 with voicing detection enabled. Pairs of glottal cycles are normalized in amplitude and scale such that analysis is based on waveform shape only,

$$\mathbf{u}_r = \uparrow_L^K \kappa_r u_D(n'), \quad (5.1)$$

where $n' \in \{n_r^c - \frac{1}{2}(n_{r+1}^c - n_{r-1}^c - 2), \dots, n_r^c + \frac{1}{2}(n_{r+1}^c - n_{r-1}^c - 2)\}$, $\uparrow_L^K$ denotes resampling factor $\frac{K}{L}$, L is the number of samples spanned by n' , $K = 2t_{max}f_s$, t_{max} corresponds to the maximum glottal period of 0.02 s and $\kappa_r = 1/||u_D(n')||_2$ is a gain factor to normalize RMS energy. The resampling process uses a polyphase filter implementation with antialiasing. The formulation ensures that each pair of cycles $\mathbf{u}_r$ contains a GCI at the centre sample while capturing as much of two neighbouring cycles as possible, which may differ slightly in duration. Normalizing cycle pairs also aids analysis by removing any dependence upon scale or amplitude. Cycle pairs from a database of speech signals form the rows of an

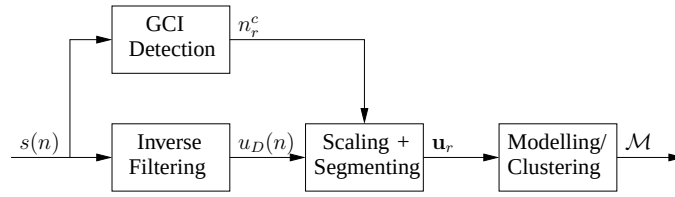


Figure 5.1: Problem formulation: training a model $\mathcal{M}$ from a speech database.

$[R \times K]$ data matrix where R is the total number of cycle pairs,

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_R]^T. \quad (5.2)$$

Model training is depicted in Figure 5.1. In the case of unvoiced speech, GCIs are inserted at a period of 10 ms and resampled in a similar manner. The segmentation into cycle pairs leads to an expectation that the rows of $\mathbf{U}$ are highly correlated for voiced speech cycles; the aim of model $\mathcal{M}$ is to describe the rows of $\mathbf{U}$ with a set of independent parameters $K' < K$. Throughout the chapter, training and test data originate from non-overlapping partitions of the APLAWD database.

5.2.1 Concatenation of Voice Source Frames for Resynthesis

Figure 5.2 shows the process of analysis/synthesis with the DDVSM framework. The approximation to the voice source signal, $\hat{u}_D(n)$, can be found by the concatenation of resampled, scaled and windowed $\hat{\mathbf{u}}_r$. Most speech processing applications assume a fixed-length frame with constant overlap of $\sim 25 - 75$ % using windows for constant-energy crossfades. Glottal-synchronous frames are generally not of uniform length so an alternative windowing scheme must be devised.

Let $N_{\text{ov}} = f_s / f_{\text{max}}$ be a fixed overlap (crossfade) time between cycles. Let $w^{\text{in}}(n)$ and $w^{\text{out}}(n)$ be the first and last N_{ov} samples of a Hanning window of length $2N_{\text{ov}}$ which will be used to achieve smooth crossfades,

$$w^{\text{out}}(n) = \begin{cases} w(n + N_{\text{ov}}) & \text{for } n = [0, 1, \dots, N_{\text{ov}} - 1] \\ 0 & \text{otherwise} \end{cases} \quad (5.3)$$

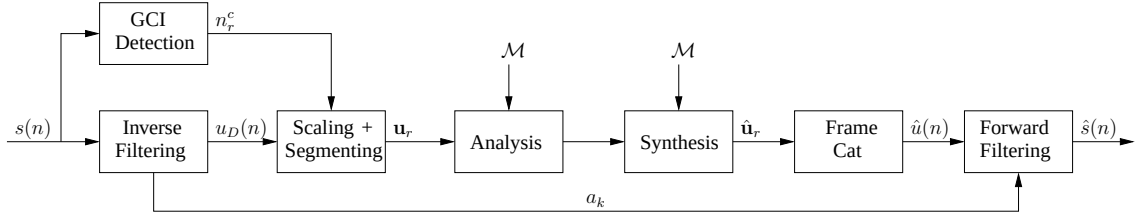


Figure 5.2: Using data-driven model $\mathcal{M}$ in the analysis-synthesis of a test signal.

$$w^{in}(n) = \begin{cases} w(n) & \text{for } n = [0, 1, \dots, N_{ov} - 1] \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

For a crossfade between two consecutive frames centered on n_r and n_{r+1} , fadeout windows $w_r^{out}(n)$ and fadein windows $w_r^{in}(n)$ are set as translated $w^{out}(n)$ and $w^{in}(n)$ respectively,

$$\begin{aligned} w_r^{out}(n) &= w^{out}(n - n_r^c - \zeta_r^{out}) \\ w_{r+1}^{in}(n) &= w^{in}(n - n_{r+1}^c + \zeta_r^{in} + N_{ov}), \end{aligned} \quad (5.5)$$

where in order to ensure constant-energy crossfades with (periodic) Hanning windows the offset parameters ζ are set as,

$$\zeta_r^{out} = \zeta_{r+1}^{in} = (n_{r+1}^c - n_r^c)/2 - N_{ov}/2. \quad (5.6)$$

The aggregate windowing function for each frame, $w_r(n) = w_r^{in}(n) + w_r^{out}(n)$, is flat topped with crossfades at ζ_r^{in} and ζ_r^{out} from the frame centre. The reconstructed voice source is

$$\hat{u}_D(n) = \sum_{r=1}^R \kappa_r w_r(n) (\downarrow_K^L \mathbf{u}_r) * \delta(n - n_r^c), \quad (5.7)$$

where $\delta(n)$ is a unit impulse function and κ restores the original energy. Note that $(\downarrow_K^L \mathbf{u}_r)$ is an noncausal signal centered at GCI r .

5.3 Feature Modelling

Feature-Domain Modelling uses existing speech feature sets to characterise voice source cycles. By assuming class separability in the features, Gaussian Mixture Modelling (GMM) [133] may be used to model feature classes and ascertain the combinations that produce highest likelihood. Let $\mathbf{f}_r$ be a vector of length F corresponding to features derived from $\mathbf{u}_r$, forming the rows of an $[R \times F]$ feature matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_R]^T$. Possible features are MFCCs [167] and PLP coefficients [168] due to their wide use in speech recognition. The likelihood of feature vector $\mathbf{f}_r$ is computed as a weighted sum of Gaussians,

$$\begin{aligned} f(\mathbf{f}_r) &= \sum_{m=1}^M p(\omega_m) f(\mathbf{f}_r | \omega_m) \\ &= \sum_{m=1}^M p(\omega_m) \frac{\exp(-\frac{1}{2}(\mathbf{f}_r - \mu_m)^T \Sigma_m^{-1} (\mathbf{f}_r - \mu_m))}{\sqrt{(2\pi)^F |\Sigma_m|}} \end{aligned} \quad (5.8)$$

where $p(\omega_m)$, μ_m and Σ_m are the weight, mean vector and diagonal covariance matrix of the m -th mixture component ω_m , $m \in \{1, 2, \dots, M\}$. The training initialization is based on K -means clustering of the data [169]. The cluster centroids produced by the K -means algorithm span the same subspace as that of the principal components and is therefore appropriate to use for this application [170]. For initialization of the GMM, mean vectors are set to the cluster means, the covariance matrices are computed from the cluster members and the initial weights are set to be the proportion of data vectors belonging to that cluster. The mixture means, variances and weights are then estimated using expectation maximization [133], terminating after 100 iterations or when increment in log likelihood falls below 0.0001.

Let $\mathbf{P}$ be an $[R \times M]$ matrix of probabilities of class ω_m given feature vector $\mathbf{f}_r$,

$$\mathbf{P} = \begin{bmatrix} p(\omega_1 | \mathbf{f}_1) & \cdots & p(\omega_M | \mathbf{f}_1) \\ \vdots & \ddots & \vdots \\ p(\omega_1 | \mathbf{f}_R) & \cdots & p(\omega_M | \mathbf{f}_R) \end{bmatrix}, \quad (5.9)$$

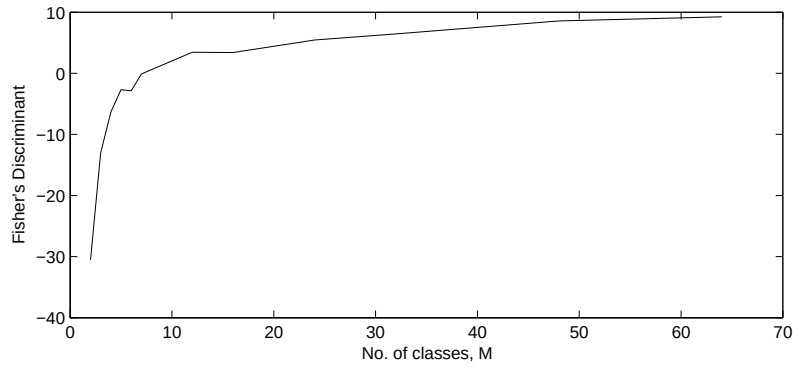


Figure 5.3: Fisher's Discriminant as a function of M classes.

where posterior probability $p(\omega_m|\mathbf{f}_r)$ is given by

$$p(\omega_m|\mathbf{f}_r) = \frac{p(\omega_m)f(\mathbf{f}_r|\omega_m)}{f(\mathbf{f}_r)}. \quad (5.10)$$

The probability that a feature vector is a member of a class may be of interest in the field of speech analysis and recognition. However, if resynthesis is to be achieved based upon $p(\omega_m|\mathbf{f}_r)$ only then an inverse is required to obtain a time-domain 'prototype' signal that is representative of the class. Features are not generally invertible so an approximation must instead be found. Before doing so, the number of classes must first be estimated.

5.3.1 Model Complexity

A suitable number of classes M is an unknown variable for which real-world data is required. Let the feature set be MFCCs, computed using 29 Mel filter banks, discarding the '0-th' and last 16 coefficients leading to the dimensionality F of $\mathbf{f}_r$ equal to 12. Fisher's Discriminant [171], F_m , measures the ratio of the intra- to inter-class variances; the higher the figure the more separated the classes. Randomly selecting half the speech samples as training data and the other half as test data, F_m was calculated, varying M from 2 to 64. Figure. 5.3 shows Fisher's Discriminant as a function of M . Asymptotic behaviour is seen beyond around $M = 16$, providing evidence for the assumption of class separability. Clustering techniques used in the remainder of this chapter assume $M = 16$.

5.3.2 Prototype Voice Source Signals

Let $\bar{\mathbf{u}}_m$ represent a time-domain *prototype* signal corresponding to the cluster ω_m , derived as a weighted average of time-domain waveforms $\mathbf{u}_r$. The weights are the probabilities $p(\omega_m|\mathbf{f}_r)$ as follows

$$\bar{\mathbf{u}}_m = \kappa_m \sum_r p(\omega_m|\mathbf{f}_r) \mathbf{u}_r, \quad (5.11)$$

where κ_m is a constant that normalizes RMS energy. The $\bar{\mathbf{u}}_m$ form the rows of the $[M \times K]$ prototype matrix,

$$\bar{\mathbf{U}} = [\bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \dots, \bar{\mathbf{u}}_M]^T = \mathbf{P}^T \mathbf{U}. \quad (5.12)$$

Alternatively, $\bar{\mathbf{u}}_m$ can be chosen from the $\mathbf{u}_r$ with maximum likelihood that it is a member of class ω_m .

$$\bar{\mathbf{u}}_m = \mathbf{u}_r \text{ s.t. } \max_i p(\omega_m|\mathbf{f}_r). \quad (5.13)$$

Many existing models of voice source waveforms include some scale-independent parameters, including the *basic shape parameter* [164], defined as $\min(u_D(n))/\max(u_D(n))$, the open quotient (OQ) [34], and the duration of the return phase [164]. In [33], a polynomial *fine* detail model describes the error between measured $u_D(n)$ and the Fant model, attributed mainly to nonlinear interaction between the glottis and vocal tract. Figure 5.4 shows 5 of the 16 classes (plots of all sixteen classes can be found in B.1). The prototypes exhibit very low noise and very little *overshoot* from LPC framing errors [172]. Variation in basic shape parameter, fine-detail ripple and duration of the open phase can also be seen. The remaining prototypes exhibit variation in all these parameters and, additionally, provide an insight into interdependencies between them.

5.3.3 Analysis-Synthesis

A test utterance can be decomposed into time-varying AR coefficients and voice source prototypes, $\bar{\mathbf{u}}_m$ as depicted in Figure 5.2. In a similar manner to signal frames for prototype training in (6.9), the test utterance is split into R overlapping frames, $\mathbf{u}_r$, where frame r

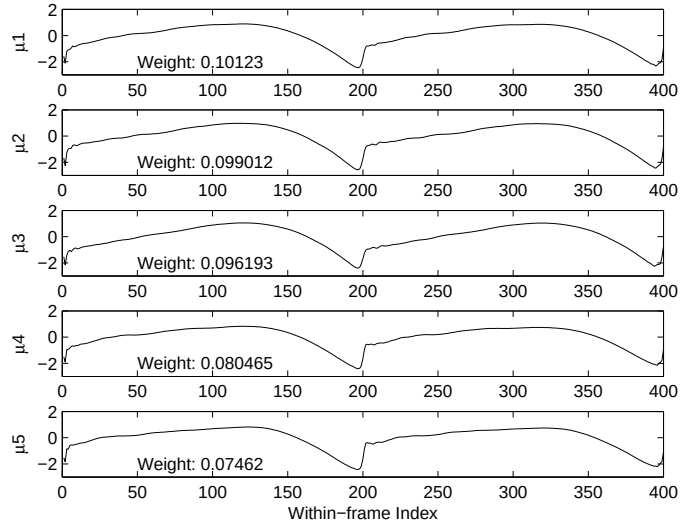


Figure 5.4: Five MFCC-derived prototype voice source waveforms in descending order of weight.

contains two cycles of voiced speech. The voice source decomposition for frame r is

$$\boldsymbol{\varphi}_r = [\varphi_{1,r}, \varphi_{2,r}, \dots, \varphi_{M,r}] = [p(\omega_1|\mathbf{u}_r), \dots, p(\omega_M|\mathbf{u}_r)]. \quad (5.14)$$

Define a set $\Gamma_r \subseteq \Gamma_{all}$, $\Gamma_{all} = \{1, \dots, M\}$. Γ_r contains the class indices that produce the highest likelihood, reducing computational complexity and bandwidth. A cycle pair of voice source signal can then be resynthesized from the prototypes, $\bar{\mathbf{u}}_m$, with the decomposition terms,

$$\hat{\mathbf{u}}_r = \sum_{m \in \Gamma_r} \varphi_{m,r} \left(\uparrow_K^L \kappa_{m,r} \bar{\mathbf{u}}_m \right), \quad (5.15)$$

where $\uparrow_K^L$ resamples $\bar{\mathbf{u}}_m$ to length L for cycle r and $\kappa_{m,r}$ is a gain factor to reproduce the same energy as the source cycle. The frames are concatenated and speech synthesized as described in Section 5.2.1.

5.4 Transform Modelling

Feature modelling is a straightforward approach to data-driven voice source modelling. An application to ABWE is presented in Section 6.4 where narrowband features are used to estimate successfully wideband speech. However, the noninvertibility of features is a significant flaw as it prevents perfect resynthesis. The averaging and weighted sums used to generate prototypes and resynthesize a signal are also problematic because they rely on perfect alignment of voice source cycles. Misalignment has the effect of smearing the GCI, attenuating the high frequency components. A final flaw is its inability to model unvoiced speech which is generally zero-mean coloured noise. Over a large number of frames, the averaging operation in the derivation of prototypes therefore sums to near-zero. A transform modelling approach circumvents the need for cycle averaging in feature modelling by finding a linear transform for the rows of $\mathbf{U}$ that is optimized in some sense. Providing an inverse-transform exists, perfect resynthesis of both voiced and unvoiced frames is possible. Furthermore, resynthesis with a restricted number of coefficients will cause a level of degradation that may be estimated from the optimization criteria.

5.4.1 Principle Component Analysis

Principal Component Analysis (PCA) is a procedure for transforming a set of correlated variables into a smaller number of uncorrelated variables called the principal components. The first component is designed to capture the largest amount of variance in the dataset as possible. Each succeeding component, which is orthogonal to all those that precede it, captures as much of the remaining variance as possible. It is beneficial as it provides a method for representing the voice source waveform as linear combination of orthogonal components.

Let $\mathbf{X}$ be an $[R \times K]$ matrix of voice source cycles with zero empirical mean,

$$\mathbf{X} = \mathbf{U} - \mathbf{h}\bar{\mathbf{u}}^T \quad (5.16)$$

where $\bar{\mathbf{u}}$ is the empirical mean of the rows of $\mathbf{U}$ and $\mathbf{h}$ is a $[R \times 1]$ vector of 1s. The $[R \times K]$

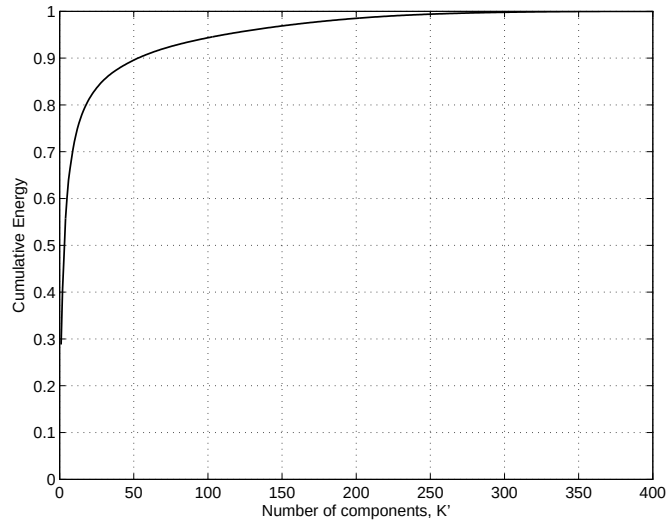


Figure 5.5: The cumulative energy of the subset of $K' < K$ principal components.

PCA transformation of $\mathbf{X}$ is given by

$$\mathbf{Z} = \mathbf{X}\mathbf{V}, \quad (5.17)$$

where $\mathbf{V}$ is a $[K \times K]$ matrix of eigenvectors of the covariance matrix $\mathbf{\Sigma} = E\{\mathbf{x}\mathbf{x}^T\}$. The columns of $\mathbf{V}$ are ordered on the eigenvalues $\lambda_1 > \lambda_2 > \dots > \lambda_K$, which are a measure of the variance captured by each eigenvector. Figure 5.5 measures the cumulative energy (variance) $\sum_{i=1}^{K'} \lambda_i / \sum_{i=1}^K \lambda_i$, contained within subset K' components for a training set of 20% of the APLAWD database. It reveals that 90% of the signal energy is contained within the first 50 components, suggesting that the intrinsic dimensionality of the voice source is significantly less than K .

The principal components in $\mathbf{V}$ are plotted in Figure 5.6. All the vectors are of unit length and are orthogonal to each other. The discontinuity due to the GCI in the centre of the $\mathbf{u}_r$ is prominent in the first 50 components. For 100 components or more the eigenvectors model noise due to aspiration and/or undermodeling of the inverse filtering process. The voice source waveform mean vector $\bar{\mathbf{u}}$ and the first four principle components $\mathbf{v}_k$ are shown in Figure 5.7. All the components model the excitation with abrupt change at the glottal closure instants. The mean vector captures the average shape of the waveform

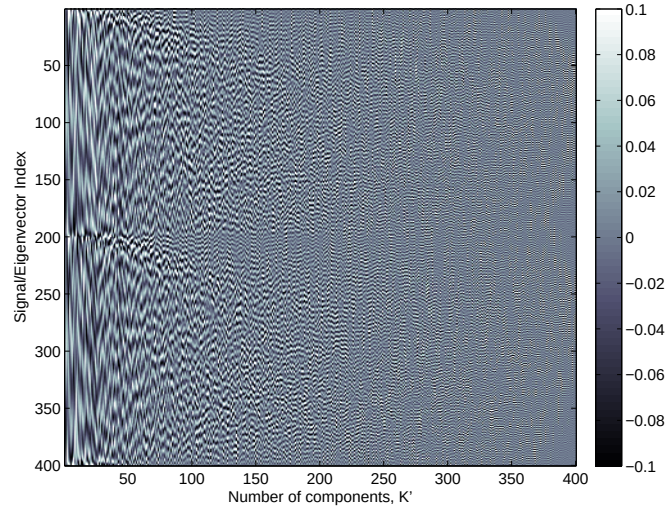


Figure 5.6: Principal components for voice source data matrix. The vectors are ordered according to significance from left to right.

whereas the first two components model the flatness in the closed phase and the slope of the opening.

Analysis-Synthesis

Resynthesis of voice source cycles from the PCA spectra is straightforward. The ordering of the columns $\mathbf{V}$ in order of eigenvalue means that the selection of the first K' columns of $\mathbf{V}$ minimizes reconstruction error by a minimum in a least-squares sense. The orthogonality of the transform is also advantageous as its inverse is the transpose of $\mathbf{V}$. Adding back the empirical mean then approximates $\mathbf{U}$ as

$$\hat{\mathbf{U}} = \hat{\mathbf{X}} + \mathbf{h}\bar{\mathbf{u}}^T \quad (5.18)$$

$$= \mathbf{Z}'\mathbf{V}'^T + \mathbf{h}\bar{\mathbf{u}}^T, \quad (5.19)$$

where $\mathbf{Z}'$ is a $[R \times K']$ matrix of containing the first K' PCA spectra and $\mathbf{V}'$ is a $[K \times K']$ matrix of the first K' PCA bases. An analysis-synthesis experiment is presented in Section 5.5.2 where Segmental Signal-to-Noise Ratio (SSNR) and Bark Spectral Distortion (BSD) are plotted as a function of PCA spectra K' .

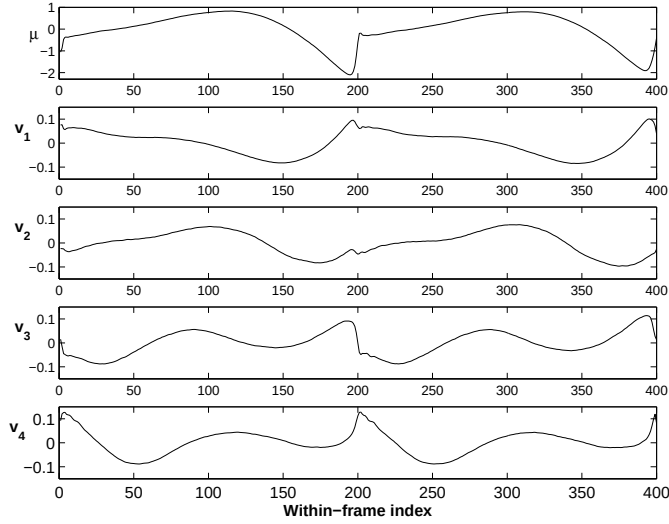


Figure 5.7: The mean vector and the first four principal components.

5.4.2 GMM Analysis on PCA Spectra

PCA provides an orthogonal basis that is optimized to capture the maximum possible variance with K' components. However, it does not take into account the class-separability of the data [170] which is shown to exist by the Fisher Ratio test in Section 5.3.1. Applying a GMM on the columns of the PCA spectra $\mathbf{Z}'$, $\mathbf{z}'_r$, reveals class separability where the likelihood function is a weighted sum of Gaussians,

$$\begin{aligned} f(\mathbf{z}'_r) &= \sum_{m=1}^M p(\omega_m) f(\mathbf{z}'_r | \omega_m) \\ &= \sum_{m=1}^M p(\omega_m) \frac{\exp(-\frac{1}{2}(\mathbf{z}'_r - \mu_m)^T \Sigma_m^{-1} (\mathbf{z}'_r - \mu_m))}{\sqrt{(2\pi)^{K'} |\Sigma_m|}}, \end{aligned} \quad (5.20)$$

where $p(\omega_m)$, μ_m and Σ_m are the weight, mean vector and covariance matrix (diagonal) of the m -th mixture component ω_m . Parameters are calculated using the same training procedure described in section 5.3. The mixture means can be resynthesized using (5.18) and interpreted as prototype voice source vectors similar to those discussed in Section 5.3.2. These are shown Figure 5.8 where the components with the 5 highest weights in the mixture are plotted (all sixteen prototypes can be found in B.2). The figure shows how different properties vary from component to component. As with the MFCC-derived prototypes

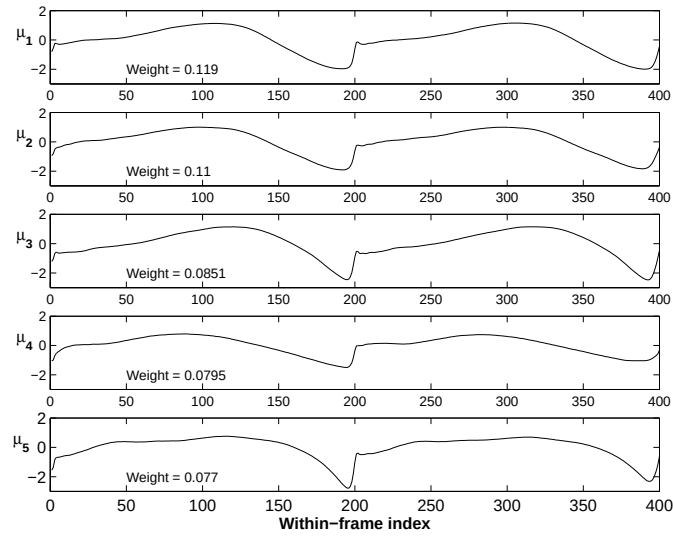


Figure 5.8: PCA/GMM Prototypes

in Figure 5.4, variation in the basic shape parameter, open quotient and opening phase ripple are seen.

Clustering of PCA spectra is discussed in a speech analysis context in Section 5.5.3 where class probabilities are plotted as a function of time. Speaker-dependence is investigated in Section 5.5.4.

5.4.3 Alternative Orthogonal Transforms

An alternative transform can be obtained by an orthonormalization of the data matrix, $\mathbf{U}$, where orthogonal bases are selected according to different optimization criteria. Such a transform is desirable as, like PCA, it is data-driven and orthogonal thus providing a compact signal representation that is easily invertible. By relaxing the least-squares constraint that exists in PCA modelling, methods accounting for features such as class separability or perceptually-weighted error might be devised. The requirement that the data matrix $\mathbf{X}$ used in PCA be zero mean necessitates the subtraction of $\bar{\mathbf{u}}$ from the rows of $\mathbf{U}$, where the empirical mean is calculated in a similar fashion to (5.11), applying uniform weights on every row of $\mathbf{U}$. It therefore suffers from the same attenuation of high-frequency components at the GCI as the prototype signals $\bar{\mathbf{u}}_m$ in feature modelling when too few

PCA spectra are used in resynthesis. A noticeable reduction in high-frequency content does not necessarily manifest itself as a large mean square error as the duration of the GCI is very short; therefore the addition of more (but not all) PCA spectra will reduce mean square error but may not bring about a perceptual improvement if the high frequency content remains missing. A method avoiding mean subtraction is therefore desirable.

The following iterative method is similar to Gram-Schmidt orthonormalization with the exception that new bases are selected according to some criteria and not in ascending order of row (or column) entry. It is suggested in [173] that a maximally sparse representation may be obtained by applying an orthonormalization process that maximises the L^2 -norm and minimises the L^1 -norm of the data. In this case a preconditioning of the matrix $\mathbf{U}$ is applied by dividing each entry by its L^1 -norm,

$$\tilde{\mathbf{u}}_r = \frac{\mathbf{u}_r}{\|\mathbf{u}_r\|_1}. \quad (5.21)$$

An iterative procedure may then be applied that selects an existing $\tilde{\mathbf{u}}_r$ according to an optimization criteria that is used as the k th basis vector $\boldsymbol{\psi}_k$; in [173] this is the frame with greatest L^2 -norm, $\tilde{\mathbf{u}}_{\hat{r}}$,

$$\boldsymbol{\psi}_k = \tilde{\mathbf{u}}_{\hat{r}_k} \text{ where } \hat{r}_k = \arg \max_r (\|\tilde{\mathbf{u}}_r\|_2), \quad (5.22)$$

although any alternative selection criteria can be used. In order to ensure that an orthogonal basis is created, the projections of $\tilde{\mathbf{u}}_r^k$ onto the new basis vector are then subtracted from the data to find residual vectors $\tilde{\mathbf{u}}_r^{k+1}$,

$$\tilde{\mathbf{u}}_r^{k+1} = \tilde{\mathbf{u}}_r^k - \frac{\tilde{\mathbf{u}}_r^{kT} \boldsymbol{\psi}_k}{\boldsymbol{\psi}_k^T \boldsymbol{\psi}_k} \boldsymbol{\psi}_k, \quad (5.23)$$

where $\tilde{\mathbf{u}}_{\hat{r}}^{k+1} = 0$ by definition. To create an orthonormal basis for $\mathbf{U}$, the k th basis vector is normalized with respect to its L^2 -norm,

$$\boldsymbol{\psi}'_k = \frac{\boldsymbol{\psi}_k}{\|\boldsymbol{\psi}_k\|_2}. \quad (5.24)$$

The process is repeated from (5.22)–(5.24) for $k = [1, 2, \dots, K]$, yielding a $[K \times K]$ trans-

form matrix $\Psi = [\psi'_1, \psi'_2, \dots, \psi'_K]^T$. Upon iteration k , all previous $\tilde{\mathbf{u}}_{\hat{r}_k}$ may be excluded as $\tilde{\mathbf{u}}_{\hat{r}_i} = 0$ for $i = [1, 2, \dots, k-1]$. The forward transform for test data is similar to (5.17),

$$\mathbf{Y} = \mathbf{U}\Psi, \quad (5.25)$$

with the corresponding *lossy* inverse after (5.18),

$$\hat{\mathbf{U}} = \mathbf{Y}'\Psi'^T, \quad (5.26)$$

where $\mathbf{Y}'$ is a $[R \times K']$ matrix containing the first K' spectra and Ψ' is a $[K \times K']$ truncated matrix of orthogonal bases. In some cases the first K' bases may not be an optimal choice. An alternative method for truncating $\mathbf{Y}$ and Ψ is to take not the first K' entries but a set of entries, whose cardinality equals K' , that maximises some optimization criterion such as minimum mean square error. Such an approach is used, for example, in image coding where the image is divided into equal-sized regions and largest K' aggregate transform spectra are transmitted and the rest discarded [174]. In the case of speech coding the number of cycles for which the aggregate spectra should be measured is a field for further research.

5.5 Experimentation

This section contains three experiments using DDVSM to demonstrate its applicability to real speech processing applications. A detailed investigation into the use of feature modelling in DDVSM is given in Section 6.4 where it is applied to artificial bandwidth extension of telephone speech. The test signals used in Sections 5.5.2 and 5.5.3 were not included in the training data.

5.5.1 An Enhanced LPC Preemphasis Filter

A description of the two-pole model of $u_G(n)$ and hence single-pole model of $u_D(n)$ was given Section 2.5. Let us assume that this model, and the corresponding single zero preemphasis filter, are over-simplified. This is discussed in [175] which states that preemphasis

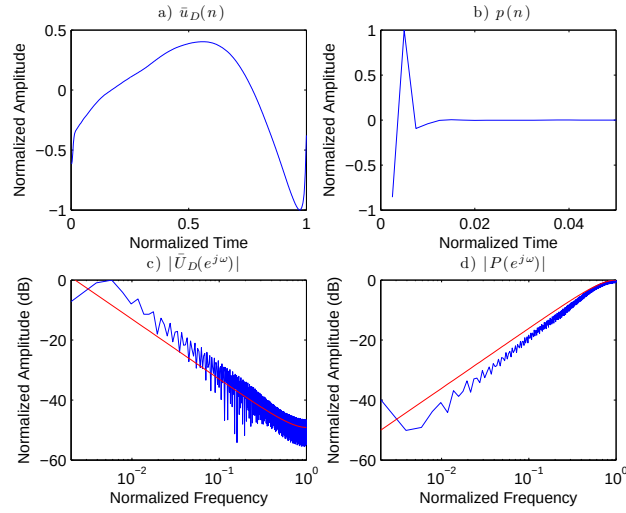


Figure 5.9: a) Time-domain excitation signal, $\bar{u}_D(n)$, b) Time-domain inverse excitation signal, $p(n)$, c) Frequency-domain excitation signal, $|\bar{U}_D(e^{j\omega})|$ and 1st order integrator, d) Frequency-domain inverse excitation signal, $|P(e^{j\omega})|$ and 1st order differentiator.

does not remove any unwanted frequency components but a) aids LPC in determining $V(z)$, particularly for higher frequency formants and b) improves the conditioning of the autocorrelation matrix. Consider instead the empirical mean excitation waveform, $\bar{u}_D(n)$, used in (5.16) in vector form. From this, an enhanced preemphasis filter, $p_{enh}(n)$, with z -transform $P_{enh}(z)$, satisfies

$$p_{enh}(n) * \bar{u}_D(n) \simeq \delta(n), \quad (5.27)$$

whose least-squares solution can be found to be

$$\mathbf{p}_{enh} = \mathbf{R}_{\bar{u}\bar{u}}^{-1} \mathbf{r}_{\bar{u}\delta} \quad (5.28)$$

where $\mathbf{p}_{enh} = [p_{enh}(0) \ p_{enh}(1) \ \dots \ p_{enh}(K-1)]$, $\mathbf{R}_{\bar{u}\bar{u}}$ is an autocorrelation matrix formed from $\bar{u}_D(n)$ and $\mathbf{r}_{\bar{u}\delta}$ is a cross-correlation vector formed from $\bar{u}_D(n)$ and $\delta(n)$.

The effect of averaging is to attenuate noise and any remaining effects of $V(z)$ not removed by inverse filtering. Figure 5.9 shows a) $\bar{u}_D(n)$, b) its least-squares inverse filter and c), d) their corresponding frequency-domain plots. In b), only the first few taps of the inverse filter are shown as it is close to a perfect differentiator with the majority

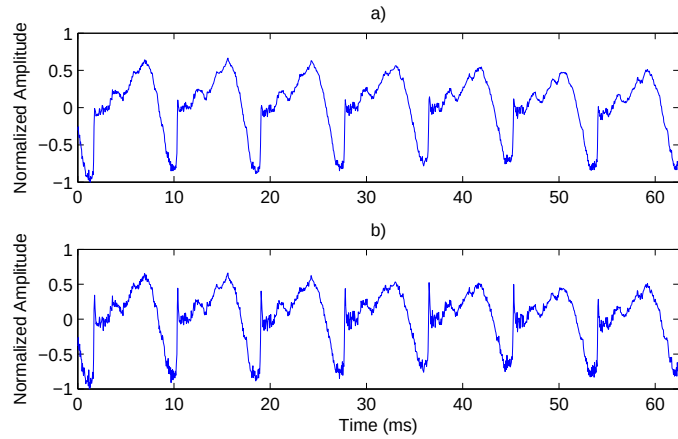


Figure 5.10: a) $u_D(n)$ derived with proposed preemphasis filter, b) $u_D(n)$ derived with standard preemphasis filter

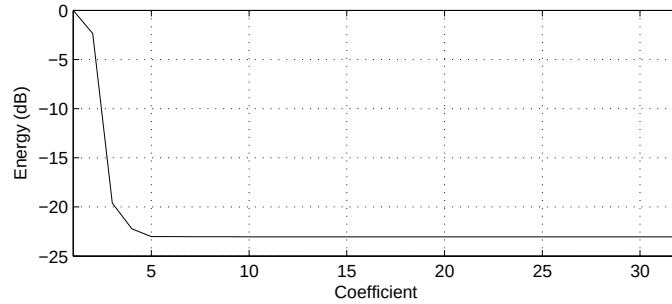


Figure 5.11: Energy decay curve for $p_{enh}(n)$.

of taps close to zero. The frequency domain plots in c) and d) show straight slopes of slightly greater gradient than the 6 dB/oct predicted by the traditional single pole model. Figure 5.10 shows a) $u_D(n)$ estimated with autocorrelation LPC using $p_{enh}(n)$ preemphasis and b) $u_D(n)$ derived with conventional preemphasis. The key improvements in (a) compared with (b) are the reduced noise during the closed phase and the reduced overshoot at the GCI caused by improved estimation of the vocal tract filter coefficients. In order to reduce computational complexity, $p_{enh}(n)$ can be truncated so that the near-zero samples are eliminated. The energy decay curve [176] in Figure 5.11 shows that most of the energy is contained within the first 5-10 coefficients. In Section 4.4, this enhanced preemphasis filter is shown to achieve an improvement in the accuracy of a GCI/GOI detection algorithm over conventional preemphasis.

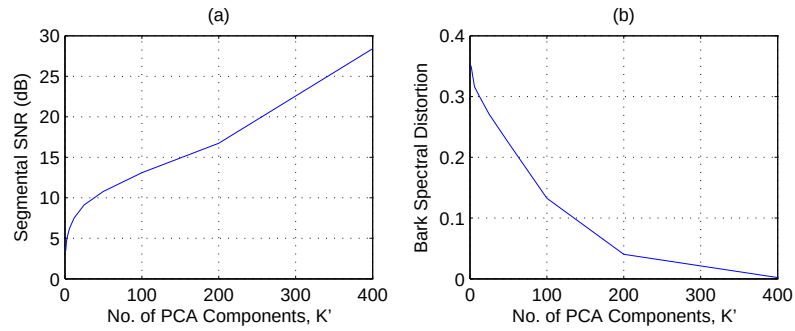


Figure 5.12: PCA Resynthesis (a) Segmental SNR and (b) Bark Spectral Distortion as a function of PCA components K' .

5.5.2 Analysis-Synthesis

A corpus of test signals from the APLAWD database [119] was decomposed with the PCA approach and resynthesized with varying numbers of components K' . SSNR and BSD are objective measures for assessing reconstruction error. The results in Figure 5.12 show (a) mean SSNR and (b) mean BSD as a function of K' . Mild degradation is observed for a reduced number of components K' , giving 12 dB SSNR for $K/8$ of the original spectra, which is evidence for its applicability to a coding scheme. The maximum SSNR of 29 dB is caused by loss of information in the resampling and concatenation process as perfect reconstruction of any given cycle can be achieved with the PCA approach. Informal listening tests give perceptually perfect reconstruction for all cases with $K' > \sim K/3$.

5.5.3 GMM Analysis

Class membership as a function of time can provide information about the dynamics of the voice source in natural conversational speech. Figure 5.13 shows (a) speech signal, (b) a hard classification, $\max_m p(\omega_m | \mathbf{z}'_r)$ and (c) soft classification $p(\omega_m | \mathbf{z}'_r)$ where black $:= (p(\omega_m | \mathbf{z}'_r) = 1)$, as a function of time for a short segment of speech. The classification is piecewise-constant; an intuitive result as the state of the glottis is relatively constant for a sustained vowel. Transition regions in the class are evidence of a significant change in the voice source signal for which the probability can be thought of as a measure of

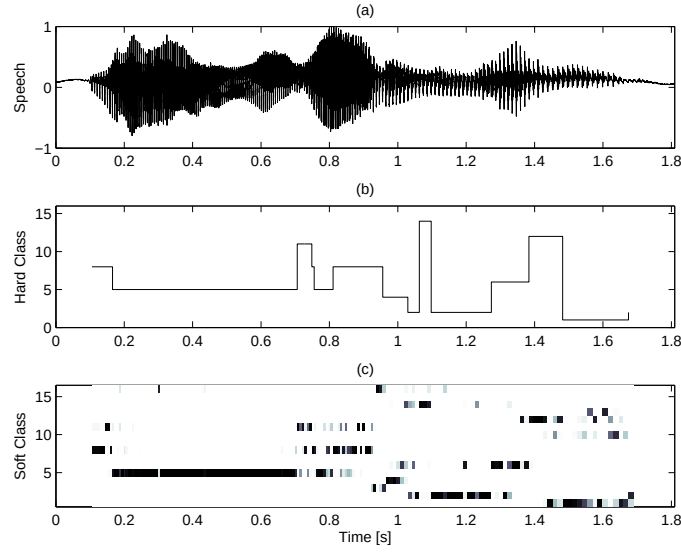


Figure 5.13: Speech signal analysis. a) Original speech signal, b) $\max p(\omega_m | \mathbf{z}'_r)$, mode-filtered with length 5, c) probability matrix $p(\omega_m | \mathbf{z}'_r)$, where black := $(p(\omega_m | \mathbf{z}'_r) = 1)$.

confidence in the class estimate. This result is of interest to both the analysis and coding communities as it represents an additional approach for compact representations of speech.

5.5.4 Speaker Dependence

Certain voice source classes can be predictors of who the speaker is. The probability of a voice source mixture component being active given a speaker is found by the expectation,

$$p(\omega_m | \xi_s) = \sum_{r=1}^{R_s} p(\mathbf{z}'_r, \omega_m | \xi_s) \quad (5.29)$$

where ξ_s is the event of speaker s talking and R_s is the number of voice source cycles from speaker s . The probability of speaker s talking given mixture component ω_m is obtained by Bayes's rule,

$$p(\xi_s | \omega_m) = \frac{p(\xi_s)p(\omega_m | \xi_s)}{p(\omega_m)}, \quad (5.30)$$

where $p(\omega_m)$ is the mixture weights and $p(\xi_s)$ is the prior probability of speaker s talking. This probability is depicted in Figure 5.14 for 10 speakers with and a GMM of 16 components. Component 3, 8, 11, 12 and 16 give proportionately higher probabilities and are therefore predictive of speakers. The other components are more equal across speakers.

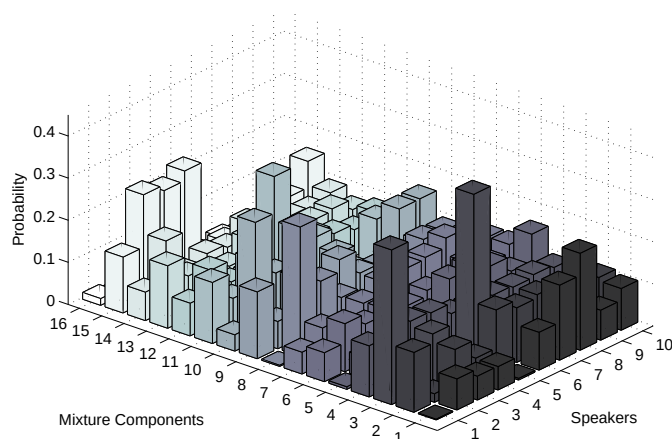


Figure 5.14: Probability of a speaker given mixture a component. Of the 16 mixture components, 5 are predictive of speaker.

5.6 Chapter Summary

A framework has been developed for data-driven models of the voice source signal. A data matrix, $\mathbf{U}$, was defined that contains pairs of voice source cycles from a speech database that are normalized in scale and amplitude so as to remove dependence upon f_0 and amplitude. Two categories of machine learning were proposed to reduce the dimensionality of $\mathbf{U}$ with a view to capturing the voice source signal with a small number of parameters.

Feature modelling is a straightforward approach to data-driven voice source modelling as it uses features commonly used in speech processing. Gaussian Mixture Models can then be applied to find clustering within the features, for which sixteen classes was shown to be a good estimate. However, it is flawed in respect of the noninvertibility of features as it prevents perfect resynthesis. The resynthesis approach employed is flawed because it relies upon perfect alignment of voice source cycles. It is also unsuitable for unvoiced speech. Transform models are elegant in that they produce an orthogonal set of bases that is optimized according to well-defined criteria. Ordering transform spectra according to their significance and reconstructing with low-order components forms the foundation of a speech compression algorithm that is suitable for both voiced and unvoiced speech. Orthogonal transforms do not necessarily take into account class separability, for which GMMs can be used to identify clustering within the data.

Practical applications have shown the applicability of the approach to real-world signals. A least-squares inverse to the average voice source signal for use in LPC preemphasis was shown in Chapter 4 to improve the accuracy of GCI detection compared with the same algorithm employing an existing voice source estimator. Analysis-synthesis with 1/8 of the original PCA spectra give a reconstruction SSNR of 12 dB. Class membership was plotted as a function of time which may be of use in coding and to understand better the dynamics of the voice source. Speaker-dependence of voice source classes showed that speaker verification might also be a future application. In the following chapter, the use of feature modelling in Artificial Bandwidth Extension is demonstrated. Data-driven voice source models therefore hold potential for areas of speech science including analysis, synthesis, coding and enhancement.

Chapter 6

Applications of Glottal-Synchronous Processing

6.1 Introduction

THE glottal-synchronous techniques presented earlier in the thesis are applied in this chapter to a series of practical problems in speech processing. In Section 6.2, a multichannel dereverberation algorithm is described that uses a spatiotemporal averaging technique driven by GCIs derived with Multichannel DYPSA (MC-DYPSA) (Section 4.5). In Section 6.3, glottal-synchronous prosodic speech manipulation using DYPSA (Section 2.8) and the Pitch-Synchronous Overlap-Add (PSOLA) Algorithm is discussed with particular reference to the importance of reliable Voiced/Unvoiced/Silence (VUS) detection. In Section 6.4, Data-Driven Voice Source Modelling (Chapter 5) is applied to Artificial Bandwidth Extension (ABWE) of telephone speech. The chapter concludes with a summary in Section 6.5.

6.2 Application 1: A Practical Multichannel Dereverberation Algorithm

Dereverberation and noise suppression play an important role in speech signal processing. As discussed in Section 2.9.1, reverberation components impair the intelligibility of a speech signal and have an adverse effect upon processing algorithms such as recognition and classification. Noise from computer fans, air ducting and other talkers can have equally undesirable consequences. A common means of attenuating these unwanted signals is beamforming, applied to an array of microphones, using the spatial diversity of room transfer functions and noise sources to attenuate the unwanted reverberation and noise components.

Beamforming is a type of *spatial averaging* which produces the greatest enhancement when the wanted components display significantly more interchannel correlation than the unwanted components. This is generally not the case for distant reflections (whose interchannel delay is low) and acoustic noise sources, so a more sophisticated algorithm is required for further enhancement. The quasi-periodicity of voiced speech can be used as a basis for *spatiotemporal averaging* [95] with the Spatiotemporal Method for Enhancement of Reverberant Speech (SMERSH). By averaging the Linear Prediction (LP) residuals over neighbouring glottal cycles from a Delay-and-Sum Beamformer (DSB), the true residual is reinforced and temporally uncorrelated reverberation and noise components are attenuated. LP synthesis with the processed residual gives a cleaner speech signal. The algorithm also uses periods of voiced speech to determine an equalisation filter [177] which performs the equivalent operation of temporal averaging for both voiced and unvoiced speech, further reducing reverberation and noise. Accurate GCI estimation from reverberant recordings is provided by the MC-DYPSA algorithm, described in Section 4.5.3. A VUS detector is further required as temporal averaging should only be applied to voiced, and therefore periodic, signals. During unvoiced speech, an equalisation filter alone is applied. The VUS detection algorithm discussed in Section 2.8.3 is applied.

Dereverberation methods can be split into three main categories: (i) beamforming

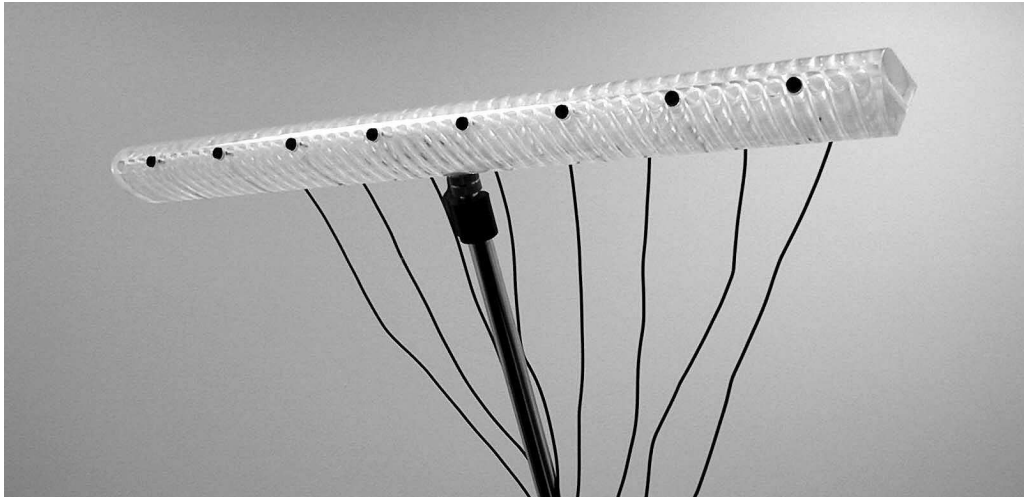


Figure 6.1: Microphone array comprising eight AKG C417 microphones placed at 5 cm intervals.

(ii) speech enhancement and (iii) blind channel estimation/equalization. Several existing algorithms are reviewed in [178]. The key contribution of this section is to combine the methods described above into a practical (online) and computationally efficient speech enhancement algorithm, which does not require knowledge of the room transfer functions and to demonstrate its applicability in real environments. The proposed method is evaluated with multichannel recordings, captured with a custom microphone array shown in Figure 6.1.

This discussion is organized as follows: Section 6.2.1 formulates the problem. Section 6.2.2 discusses the algorithm in detail. Test results are presented in Section 6.2.3.

6.2.1 Problem Formulation

Consider a speech signal $s(n)$ produced in a reverberant environment, received by an array of M microphones, through a channel $h_m(n)$ from the source to microphone m . The received signal at microphone m is $x_m(n) = h_m(n) * s(n)$, where $*$ denotes linear convolution. Given $x_m(n)$, $m = 1, 2, \dots, M$, the aim is to estimate an enhanced speech signal, $\hat{s}(n)$.

LP analysis [50], introduced in Section 2.6.1, describes a speech signal as a linear

combination of p past samples, such that

$$s(n) = \sum_{k=1}^p a_k s(n-k) + e(n) \quad (6.1)$$

where a_k are the clean LP coefficients and $e(n)$ is the clean LP residual. Similarly, LP analysis can be applied to each microphone output

$$x_m(n) = \sum_{k=1}^p b_{m,k} x_m(n-k) + e_m(n) \quad (6.2)$$

where $b_{m,k}$ are the LP coefficients for channel m and $e_m(n)$ is the corresponding LP residual. A single set of best-fit LP coefficients, b_k , may be found with multichannel LPC analysis that closely match a_k . This analysis is investigated in detail in [179].

Given $e_m(n)$, an enhanced LP residual, $\hat{e}(n)$, need be obtained such that reverberation components are reduced and the speech enhanced. The following section describes a spatiotemporal averaging technique for finding $\hat{e}(n)$, with which an enhanced speech signal can be resynthesized by LP synthesis

$$\hat{s}(n) = \sum_{k=1}^p b_k \hat{s}(n-k) + \hat{e}(n). \quad (6.3)$$

6.2.2 Proposed Approach

The practical implementation of SMERSH comprises four parts: Time Delay of Arrival (TDoA) estimation with Generalized Cross-Correlation PHase Transform (GCC-PHAT), VUS detection, GCI detection with MC-DYPSA and spatiotemporal averaging.

TDoA Estimation

Both MC-DYPSA and spatiotemporal averaging rely on the correct inter-channel time alignment to maximise the correlation of the direct-path signal across channels. GCC-PHAT [97], described in Section 2.9.1, is a simple and sufficiently accurate method for the estimation of delay between two channels from moderately reverberant speech signals.

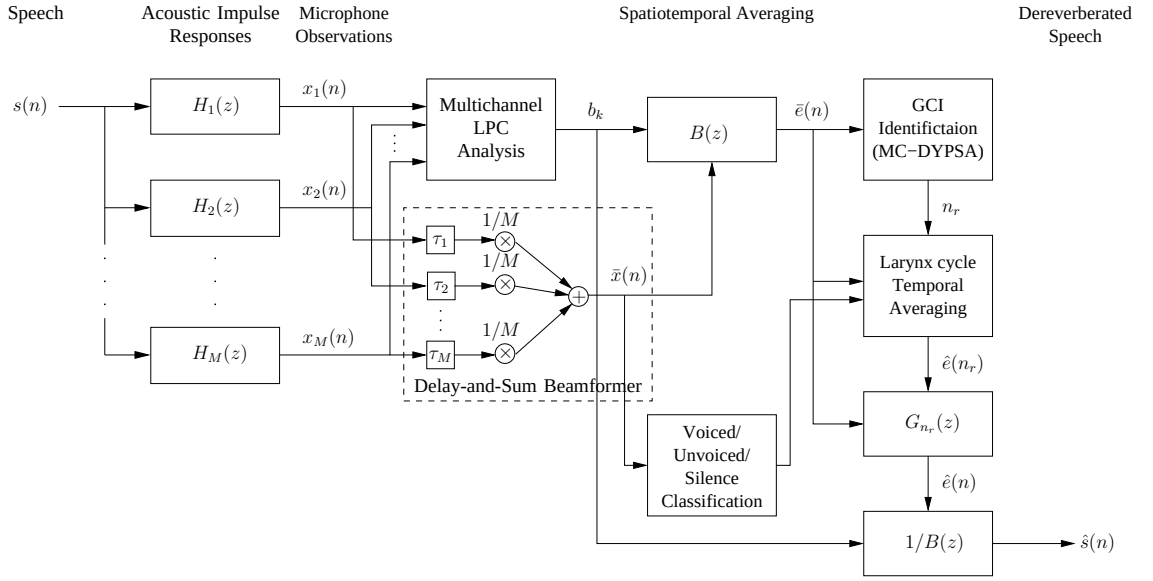


Figure 6.2: System diagram of the multichannel dereverberation algorithm.

Untrained VUS Detection

An untrained VUS detector is used, where Atal feature vectors defined Section 2.8.3 are clustered using an unsupervised EM algorithm [133]. The three clusters are labelled as silence, unvoiced and voiced according to their mean vectors and variances. The unvoiced cluster is chosen to be the one with an autocorrelation coefficient closest to zero mean and 0.5 variance. Of the remaining two clusters, the one with greatest energy is chosen to be voiced. Every frame is evaluated under each of the three Gaussians and classified according to which cluster produces the highest likelihood. Examples of this untrained VUS detection can be found in [180] and [181].

Spatial Averaging

VUS detection is performed on a speech signal which has been processed with a DSB. The output of the DSB, $\bar{x}(n)$, is found by

$$\bar{x}(n) = \frac{1}{M} \sum_{m=1}^M x_m(n - \tau_m) \quad (6.4)$$

where τ_m is a delay to compensate for the propagation time of the source to channel m .

Temporal Averaging

The DSB prediction residual, $\bar{e}(n)$, found by inverse filtering $\bar{x}(n)$ with b_k [50], contains peaks due to GCIs and many spurious peaks due to reverberation and noise. Spurious peaks are uncorrelated among consecutive glottal cycles. Conversely, the main features of prediction residuals from clean speech vary little between neighbouring cycles because of the quasi-stationarity of voiced speech. Performing a weighted average of $\mathcal{I}$ neighbouring residuals from glottal cycles of length $L = n_{r+1}^c - n_r^c + 1$ of noisy, reverberant speech reinforces the clean speech excitation and attenuates the uncorrelated spurious peaks using

$$\hat{\mathbf{e}}_r = (\mathbf{I} - \mathbf{W})\bar{\mathbf{e}}_r + \frac{1}{2\mathcal{I}} \sum_{i=-\mathcal{I}}^{\mathcal{I}} \mathbf{W}\bar{\mathbf{e}}_{r+i} \quad (6.5)$$

where $\bar{\mathbf{e}}_r = [\bar{e}(n_r) \ \bar{e}(n_r + 1) \ \dots \ \bar{e}(n_r + L - 1)]^T$ is the r th glottal cycle at the output of the DSB with GCIs at time n_r , $\hat{\mathbf{e}}_r = [\hat{e}(n_r) \ \hat{e}(n_r + 1) \ \dots \ \hat{e}(n_r + L - 1)]^T$ is the r th glottal cycle of the enhanced residual and $\mathbf{I}$ is the identity matrix. $\mathbf{W} = \text{diag}\{\omega_0 \ \omega_1 \ \dots \ \omega_{L-1}\}$ is a diagonal weighting matrix to exclude excitation at the GCIs based on the Tukey window [182]. The windowing helps when GCIs are poorly estimated as noticeable artefacts can be produced when high-energy excitations, located at the GCIs, are averaged between two misaligned cycles. By averaging $\bar{\mathbf{e}}_r$ with the high-energy GCI excitation excluded, errors caused by misalignment are perceptually reduced.

This process can only be applied to segments of voiced speech, leaving reverberation components unaffected on unvoiced speech and silence. Furthermore, in the case of an erroneous GCI, the algorithm will produce incorrect results. To improve robustness, an L_g -tap equalisation filter $\mathbf{g} = [g_0 \ g_1 \ \dots \ g_{L_g-1}]^T$ for the r th glottal cycle is defined which performs the equivalent operation of temporal averaging. A least squares estimate of $\mathbf{g}$ is found from $\hat{\mathbf{g}} = \arg \min_{\mathbf{g}} \|\mathbf{g}^T \bar{\mathbf{e}}_r - \hat{\mathbf{e}}_r\|_2^2$ whose least-squares solution can be found to be

$$\hat{\mathbf{g}} = \mathbf{R}_{\bar{\mathbf{e}}\bar{\mathbf{e}}}^{-1} \mathbf{r}_{\bar{\mathbf{e}}\hat{\mathbf{e}}} \quad (6.6)$$

where $\mathbf{R}_{\bar{\mathbf{e}}\bar{\mathbf{e}}}$ is an autocorrelation matrix formed from $\bar{\mathbf{e}}_r$ and $\mathbf{r}_{\bar{\mathbf{e}}\hat{\mathbf{e}}}$ is a cross-correlation vector

formed from $\bar{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_r$. The filter in (6.6) is used to update a slowly varying filter

$$\hat{\mathbf{g}}(n_r) = \gamma \hat{\mathbf{g}}(n_{r-1}) + (1 - \gamma) \hat{\mathbf{g}}_r \quad (6.7)$$

where $0 \leq \gamma \leq 1$ is a forgetting factor with typical values in the range $\{0.1-0.3\}$, initialised to $\hat{\mathbf{g}}(0) = [1 \ 0 \ \dots \ 0]^T$. It is updated only during voiced speech, with the last iteration used for periods of unvoiced speech or silence.

6.2.3 Evaluation

The microphone array shown in Figure 6.1, consisting of eight AKG C417 microphones spaced linearly at 5 cm intervals, was placed in a 3.3x2.9x2.9 m room with reverberation time (T_{60}) of 0.3 s. A channel estimation was made for each microphone using the Maximum Length Sequence (MLS) method [4]. Utterances of the sentence, “George made the girl measure a good blue vase,” by five male and five female talkers were taken from the APLAWD database [119] and played through a loudspeaker at distances 0.5 to 2 m from the microphone array.

The MLS-derived channel estimates were truncated to determine a direct-path impulse response, $h_d(n)$, which was convolved with the clean speech signal to align the unprocessed and processed signals, denoted $s'(n) = h_d(n) * s(n)$. Recording and channel alignment were made at a sampling frequency of $f_s = 48$ kHz. The remainder of the processing was performed at $f_s = 16$ kHz with the samples high-pass filtered at 100 Hz. The recorded, DSB and spatiotemporal averaged speech samples were evaluated against $s'(n)$ using the segmental Signal-to-Noise Ratio (SNR) [178] and Bark Spectral Distortion (BSD) [114] using 30 ms frames with 50% overlap. The definition of *noise* in this case is the combination of both reverberation and background noise.

Results and Discussion

The segmental SNR results, averaged over all ten talkers in APLAWD, are shown in Figure 6.3 for (a) reverberant speech at the microphone closest to the talker, (b) DSB speech

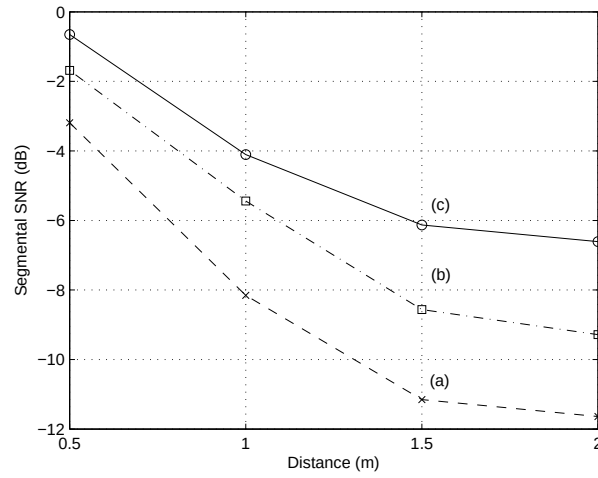


Figure 6.3: Segmental SNR vs. distance for (a) reverberant, (b) DSB processed and (c) Spatiotemporal averaged speech.

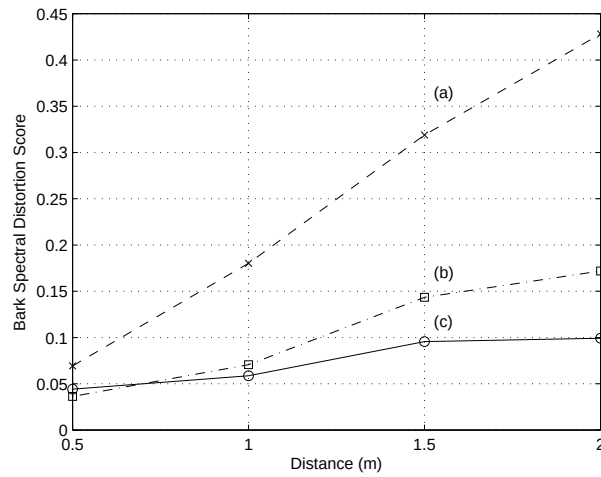


Figure 6.4: BSD vs. distance for (a) reverberant, (b) DSB processed and (c) Spatiotemporal averaged speech.

and (c) spatiotemporal averaged speech. Corresponding BSD results are shown in Figure 6.4. Reverberation and noise reduction of up to 5.0 dB and 0.33 in BSD score are observed at a distance of 2 m, corresponding to 2.7 dB and 0.07 over the DSB. Perceptually, reverberation effects are reduced and the talker appears to be closer to the microphone. The results show a strong correlation with the simulations in [177]. Examples of clean and processed samples can be found at [183].

6.3 Application 2: Time Scale Modification of Speech

Speech time scale modification is a process which alters the length of a segment of speech without significantly affecting its pitch or formant structure. It has many uses, including time scale compression for fast scanning of recorded voicemail messages [98] and time scale expansion for improving the intelligibility of fast or degraded speech in forensic applications. A combination of compression and expansion may also find uses in the synchronization of audio to lip movements in motion video. Real-world applications of time scale modification have, however, been limited due to the presence of unwanted artefacts in existing approaches. This section presents a new approach that reduces many common artefacts and provides fast and perceptually superior results.

During voiced speech, the pseudo-periodicity of the waveform naturally lends itself to time scale modification as complete glottal cycles may be removed or repeated depending upon whether a compression or expansion of signal duration is desired. Providing the periods are accurately known and cycles are concatenated in such a way that pitch periods are faithfully reproduced, good time scale modification can be achieved. However, during periods of unvoiced speech, voiced fricatives, plosives or boundaries of voiced speech, no such periodicity exists, though most algorithms still apply uniform time scale modification to the entire speech signal. These segments will be referred to hereon as *unvoiced and transition* (UT) segments. The resulting artefacts in UT segments, caused by algorithms such as the following, diminish the quality of the processed speech.

Existing approaches for concatenating periods of voiced speech for time scale/pitch modification include the PSOLA method [99] and specifically time-domain PSOLA (TD-PSOLA) which performs well provided a) pitch periods are accurately known and b) high quality time scale (but not pitch) modification is required. Other approaches include sinusoidal-based [100], LP residual-based (LP-PSOLA) [50,101], waveform similarity-based (WSOLA) [102] and phase vocoders [103], which address the cases when one or more of these constraints are unfeasible, at the cost of added complexity. More recent approaches address the issue of UT segments [105–109]. In the literature, effort has been made to apply different levels of duration modification for different segments with positive results

but little work has been done to optimize these parameters. The studies generally conclude that the most perceptually significant artefacts are those arising from the repetition of UT segments, for which the fast and accurate detection is key in the proposed algorithm. The approach also differs in the use of the DYPSA algorithm [77] to quickly and reliably find GCIs to use as pitch markers.

The strategy is to address the problem of modification during segments of little or no periodicity by employing a VUS detector, from which UT segments are derived. It assumes that the duration of most UT segments is independent of speech rate [104, 105] and does not apply modification to them. During voiced segments, DYPSA provides GCIs which are used as pitch markers. During silence, the algorithm places pseudo-pitch markers every 10 ms. Cycles are then concatenated using PSOLA, ensuring that pitch periods are faithfully reproduced using the approach in [101]. The result is a practical, fast and reliable method for time scale modification that is novel in a) the use of DYPSA to find pitch markers and b) the use of a Gaussian Mixture-based classifier to find UTs. Subjective testing has shown that the proposed method gives significantly greater mean opinion scores than an equivalent method which performs uniform processing on the entire speech signal.

This discussion is organised as follows: Section 6.3.1 formulates the problem with a set of examples. Section 6.3.2 describes the proposed approach. Results and discussion of subjective tests are presented in Section 6.3.3.

6.3.1 Problem Formulation

Compression or expansion of speech time scale involves removing or repeating cycles as required, as shown in Figure 6.5. Using a Hamming window on each pitch period, a predefined overlap and an appropriate weighting for energy normalisation, pitch periods may be cross-faded to form a resynthesised signal,

$$y(l) = \frac{s(n_r^c + l - \tilde{n}_i^c)w((l - \tilde{n}_i^c)/k) + s(n_{r+1}^c + l - \tilde{n}_{i+1}^c)w((l - \tilde{n}_{i+1}^c)/k)}{\sqrt{w^2((l - \tilde{n}_i^c)/k) + w^2((l - \tilde{n}_{i+1}^c)/k)}} \quad (6.8)$$

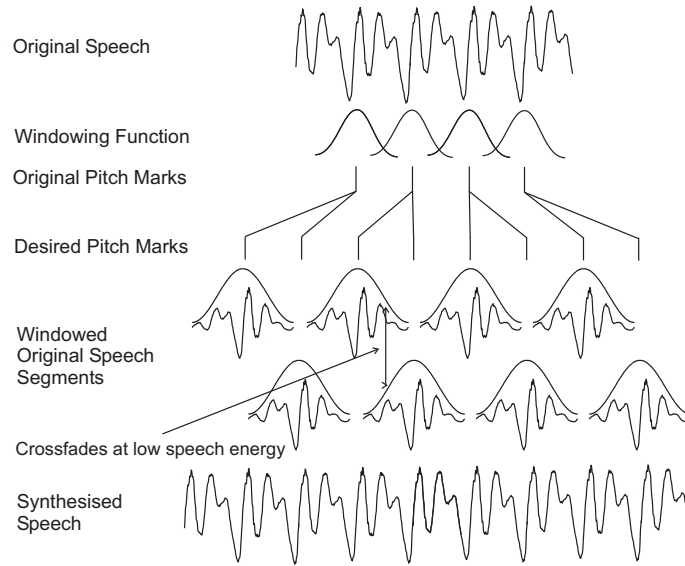


Figure 6.5: Concatenation of pitch periods (after [184]). Periodicity is identified (in this case at the instants of glottal closure) and individual periods are multiplied with a Hamming window. Periods are repeated or removed as necessary and then aligned and normalised to form a new synthesised signal with modified time scale. The use of GCIs as pitch markers ensures that crossfades take place during regions of low speech energy.

where $s(n)$ is the input speech signal, $k = \min(\tilde{n}_{i+1}^c - \tilde{n}_i^c, n_{r+1}^c - n_r^c)$, $w(x) = \frac{1}{2}(1 + \cos(\pi x))$ and $\tilde{n}_i^c$ are time-stretched pitch markers. Pitch markers must be pitch-synchronous, but by identifying GCIs, it is guaranteed that in addition to being pitch-synchronous, crossfades take place where there is low speech energy. Such an approach can give good time scale modification during periods of voiced speech providing the GCIs are accurate. A *phasiness* property often accompanies poor GCI estimation in time scale modification [103].

However, Figure 6.6 shows that although the PSOLA approach works well during voiced speech, processing with pseudo pitch periods placed every 10 ms during unvoiced sounds can cause periodic components it did not originally possess when the time scale is stretched. This gives a very unnatural-sounding result that diminishes the overall quality of the processed speech. In the case of a fast-spoken sentence which is to be slowed down, there is a greater ratio of the duration of unvoiced to the duration of voiced speech, as the duration of many unvoiced sounds has been found to be largely independent of talking rate [104, 105]. The aforementioned artefacts will therefore be worse in the case

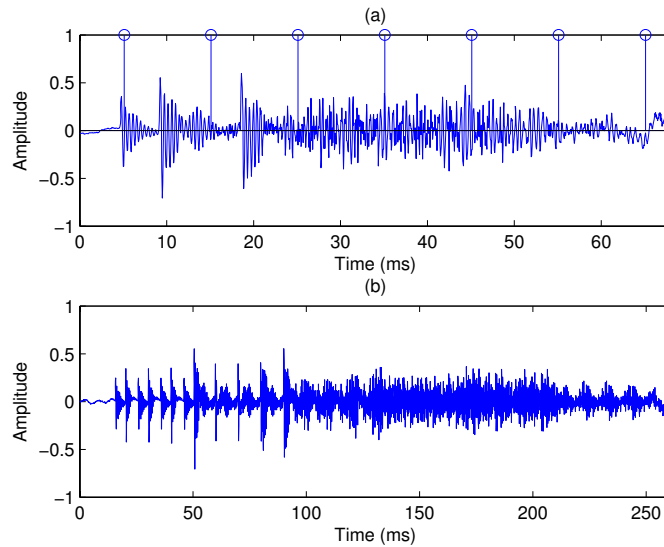


Figure 6.6: The effect of treating unvoiced sounds as periodic in time expansion. Speech signal (a) is an utterance of the phoneme /tʃ/ (containing both impulsive events and turbulent noise), with pseudo pitch markers placed every 10 ms. A time expansion of four times is shown in (b) which contains many additional harmonic components.

of time scale expansion on fast-spoken speech. When compressing the time scale of a speech signal, a uniform approach can cause short (but important) sections of speech to be removed altogether and significantly impair intelligibility. Figure 6.7 gives an example where a plosive is lost. These problems may be addressed if UT segments are isolated and left unchanged, allowing only voiced and silent periods to be modified. The assumption that many unvoiced sounds are either weakly proportional to talking rate or are entirely independent, is mentioned in [104,105] and backed up by subjective testing in Section 6.3.3.

6.3.2 Proposed Approach

Speech utterances for VUS model training were recorded by three talkers, of combined duration 100 s, under the same conditions as those used in the experiments presented in Section 6.3.3. They were labelled as $\omega \in \{\mathcal{V}, \mathcal{U}, \mathcal{S}\}$ by hand then excluded from use in the subjective test set. The data was used to train the VUS detector, described in Section 2.8.3, based on feature vectors derived from 20 ms frames of speech. Each class is modelled by a multivariate full covariance Gaussian distribution, whose parameters are derived from labelled training data. During voiced segments, DYPSA is used for GCI detection. The

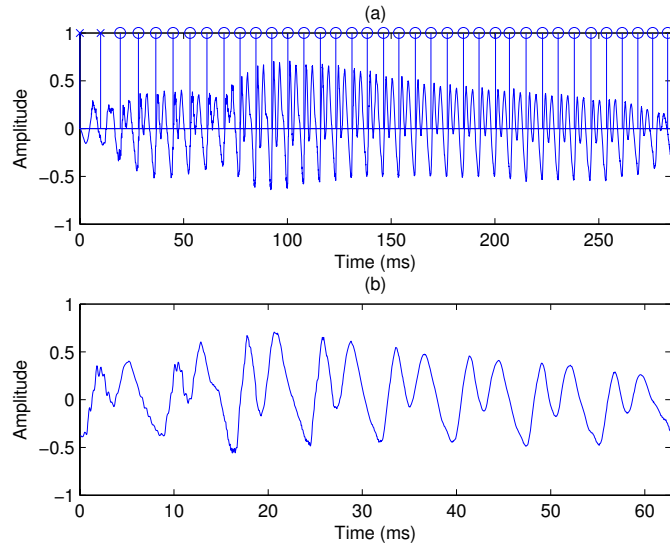


Figure 6.7: The effect of treating unvoiced sounds as periodic in time compression. Speech signal (a) is an utterance of the word /*blu*/ (blue), with pseudo pitch periods marked ‘×’ and GCIs marked ‘o’. A time compression of four times is shown in (b) which sounds closer to /*lu*/.

use of DYPSA for this purpose – which relies heavily on waveform similarity – is loosely related to the WSOLA technique [102], with the notable exception that DYPSA operates on a set of candidates derived from the LPC residual with the Group Delay method. During silence, pseudo pitch markers are placed every 10 ms. Time scale modification is then applied to the marked voiced and silent segments; UT regions are left unprocessed. The approach for determining the new pitch markers $\tilde{n}_i^c$ is based upon the method in [101].

Determination of UT, Voiced and Silent Segments

The VUS detector provides a set of probabilities for voiced, unvoiced and silence as shown in Figure 6.8. Voiced segments, $\mathcal{V}$, are identified by applying a Schmitt Trigger operator S^+ to the voiced probability, $S^+(P\{\mathcal{V}|\mathbf{x}_i\})$. Transition segments, $\mathcal{T}$, are derived by identifying the boundaries of $\mathcal{V}$ and flagging a segment 10 ms before and after the boundary. Unvoiced segments, $\mathcal{U}$, are identified by applying a Schmitt Trigger to the unvoiced probability, $S^+(P\{\mathcal{U}|\mathbf{x}_i\})$ (with upper and lower thresholds empirically set at $\{0.25, 0.75\}$) and extending in time scale by 2 ms at the boundaries. The UT segment is the union of $\mathcal{U}$ and $\mathcal{T}$, $UT = \mathcal{U} \cup \mathcal{T}$ and all remaining segments are flagged as silence, $\mathcal{S}$.

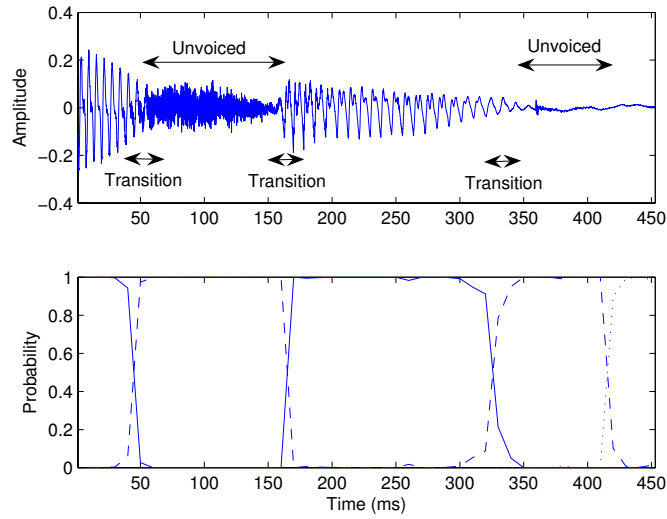


Figure 6.8: Normalised probabilities for $\omega \in \{\mathcal{V}, \mathcal{U}, \mathcal{S}\}$. Solid: voiced, dashed: unvoiced, dotted: silence.

6.3.3 Evaluation

A subjective test was performed to determine the mean opinions of the speech quality produced by two time scale modification algorithms. Both algorithms applied concatenative synthesis directly on speech recordings. Algorithm 1 performed uniform time scale modification on the entire signal and Algorithm 2 is our proposed algorithm which uses UT detection to perform time scale modification during voiced speech and silence only.

The recording apparatus comprised an AKG C480 microphone connected to an RME Fireface 800 audio interface. Subjective testing samples were played back through the same interface connected to a pair of Sennheiser HD650 headphones. Three talkers (two male, one female) were placed in an anechoic chamber and recorded speaking five phonetically balanced sentences at what they considered to be a *normal* speaking rate and a *fast* speaking rate (approximately 0.5–0.75 the duration of normal). The texts were taken from the APLAWD and TIMIT databases [119, 185]. The recorded speech was free from background noise, reverberation or any significant distortion.

An ITU-T P800 [117] double-blind controlled test was employed where 30 test subjects were each played 60 random combinations of talker {1-3}, sentence {1-5}, talking rate {*normal*, *fast*}, algorithm {*no UT detection*, *UT detection*} and time scale modifi-

cation rate $\{0.25, 0.5, \dots, 2.75, 3\}$. A modification rate of 1 implies no processing was undertaken and the subjects were not aware of what samples they were listening to, nor were they aware of how many algorithms had been employed. Sentences recorded *fast* only had time-scale expansion, so that the modification rate was always greater than 1. The test subjects were asked to give overall opinion scores in the range $\{1-5\}$, paying attention to intelligibility, prosody and artefacts. Calibrated examples were given before the test was undertaken, defined as: 1=Unsatisfactory, 2=Poor, 3=Fair, 4=Good, 5=Excellent.

Results and Discussion

Figure 6.9 shows Mean Opinion Scores (MOSs) as a function of time scale modification rate with their corresponding confidence intervals in Figure 6.10. Mean MOS scores are shown in Table 6.1. The results show that UT detection is preferred by the listeners, particularly at larger modification rates. However, at the lowest rates, 0.75 and 1.25, there may be evidence to suggest that UT detection is unnecessary. Informal listening has shown that although artefacts are reduced in the UT case, the flow is slightly interrupted so there may be a preference for smooth flow over artefacts for low levels of modification. The greatest difference in opinions occurs at a rate of 0.25 on normal speech, where some subjects described the non-UT method as *garbled* and the UT as *unnatural but intelligible* during informal listening tests. This would suggest that in the case of extreme speeding up, a listener prefers to preserve intelligibility at the cost of impairing natural flow. The control samples show the highest MOS, though it is reduced by about 0.5 for *fast* talking rate compared with *normal*; similar scores are seen for normal speech modified by 0.5-0.75. This is evidence that intelligibility is preferred over the presence of artefacts at large deviations from normal; if this were not the case then the MOS for unmodified speech would be similar regardless of the original talking rate.

Now that it has been established that segmented time scale modification is a worthwhile pursuit, an extension of this method is the discrimination between different types of speech in addition to UT detection, then applying different amounts of stretching or compression based upon training data. This is mentioned in [106] where vowels are de-

tected as a subset of voiced speech, but other cases such as inter-phoneme and inter-word pauses, stressed phonemes etc. may also vary differently with talking rate in natural speech. Examples of time scale-modified speech can be found at [183].

Table 6.1: Mean MOS scores for timescale-modified speech

	Normal Speed	Fast Speed	Mean
UT detection, μ_{UT}	3.54	3.18	3.36
No UT detection, μ_{NUT}	2.90	2.60	2.75
$\mu_{UT} - \mu_{NUT}$	0.65	0.57	0.61

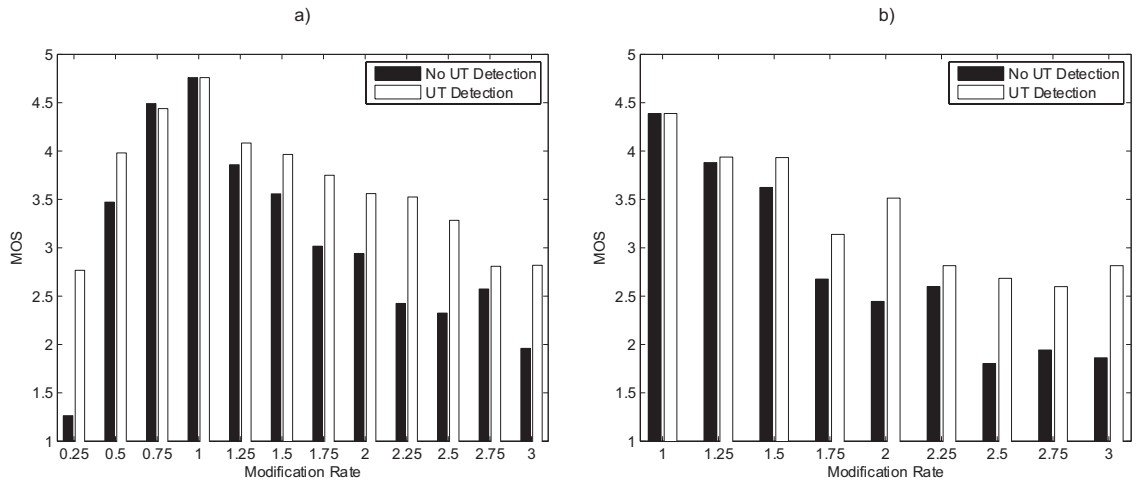


Figure 6.9: MOS scores as a function of time scale modification for a) normal original talking rate and b) fast original talking rate.

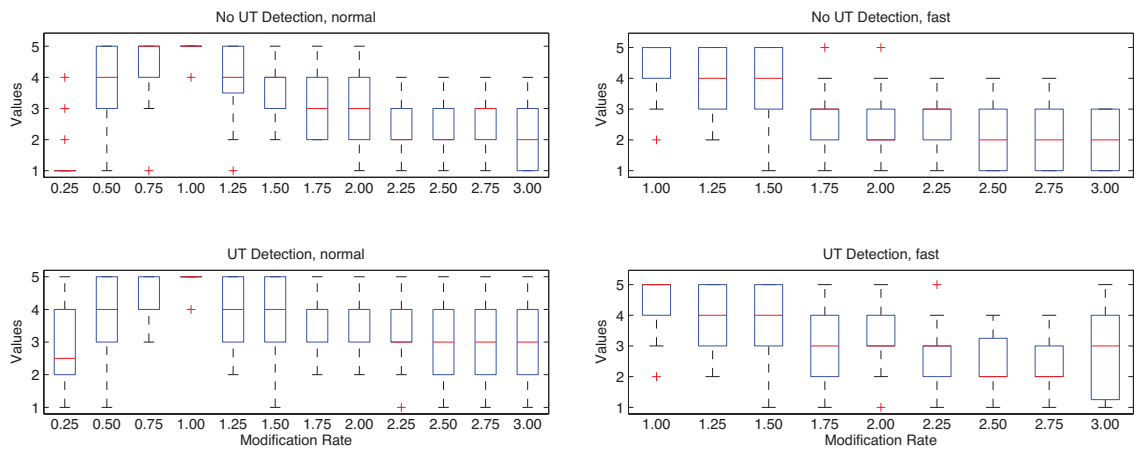


Figure 6.10: MOS confidence intervals as a function of time scale modification.

6.4 Application 3: Artificial Bandwidth Extension of Telephone Speech

The audio bandwidth of 300 Hz – 3.4 -kHz which is used in today’s fixed and mobile communication systems is comparable to that of early-day analogue telephony. When digital standards were first established, a common audio bandwidth facilitated interoperability between the analogue and digital domains. There has since been motivation within the telecommunications industry to introduce wideband telephony which can deliver high-quality speech with an audio bandwidth of 50 Hz – 7 kHz to end-user terminals. Figure 6.11 shows the spectrum of (a) unvoiced speech (/s/) and (b) voiced speech (/a/) with overlaid narrowband and wideband bandwidths. In the narrowband, unvoiced speech generally lacks high frequency energy whereas voiced speech lacks low frequency energy.

Both narrowband and wideband systems are expected to co-exist for a long time, requiring measures to ensure interoperability between narrowband and wideband telephones. This coexistence poses two main challenges: (a) efficient transcoding between narrowband and wideband signals, and (b) speech bandwidth extension to improve the quality of narrowband speech received on wideband terminals. The former can be addressed by hierarchical coding where a standard narrowband bitstream is augmented with side information to extend the audio bandwidth [46]. This approach is termed Bandwidth Extension with Side Information. Transcoding is then straightforward as the side information be either included or discarded as required [46]. In the latter case, the so-called extension bands (50 – 300 Hz and 3.4 – 7 kHz) are instead *estimated* from the narrowband speech only. This is referred to as Artificial Bandwidth Extension (ABWE).

Most ABWE methods use the source-filter model of speech production to estimate wideband spectral and temporal envelopes independently of the source signal. Appropriate techniques to blindly estimate these envelopes include codebook mapping [110], piece-wise linear mapping [111] and Bayesian methods based on Gaussian Mixture Models (GMMs) [112] or Hidden Markov Models (HMMs) [113]. Although the existing methods can already deliver superior quality compared to narrowband speech, many ABWE

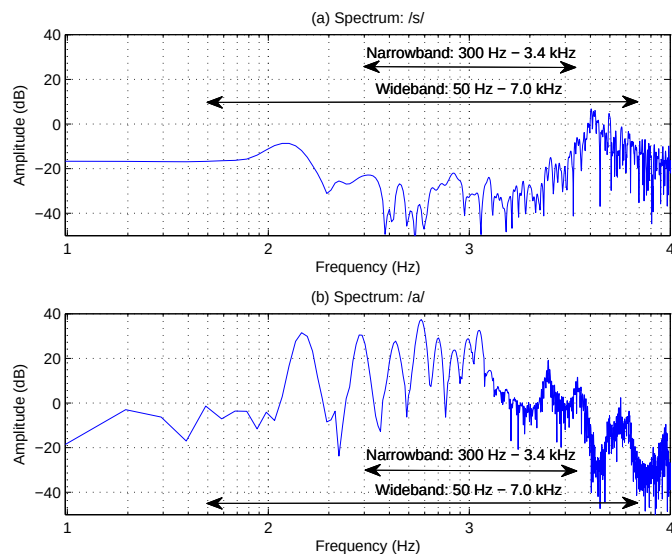


Figure 6.11: Speech Spectra for (a) /s/ (unvoiced) and (b) /a/ (voiced) with narrowband and wideband bandwidths.

algorithms employ relatively crude methods to extend the source signal. For ABWE towards high frequencies (3.4 – 7 kHz) there is evidence that the quality of the enhanced speech mainly depends on a precise estimate of the spectral envelope while the source signal extension is less important [5]. However, if low audio frequencies (50 – 300 Hz) are also to be recovered from narrowband speech, existing source extension methods usually fail to produce a signal of sufficient quality, in particular for voiced speech segments. Typical artefacts include a roughness caused by low-frequency random noise that is modulated by the speech amplitude, or a buzziness caused by incorrectly shaped or incorrectly placed glottal pulses, depending upon the method employed. Such artefacts render the bandwidth-extended speech unnatural and can mask any perceived improvement in quality. For this reason, existing ABWE approaches often avoid lowband extension altogether.

This section presents a novel method for the extension of narrowband source signals based on an existing spectral mirroring technique and Data-Driven Voice Source Modelling (DDVSM) [145], described in Chapter 5, employing GMMs to establish an explicit mapping between narrowband source features and the wideband source signal. Using an existing ABWE framework [113] that applies HMM-based Bayesian estimation of spectral and temporal envelopes [46], missing frequency content in both high and low bands is

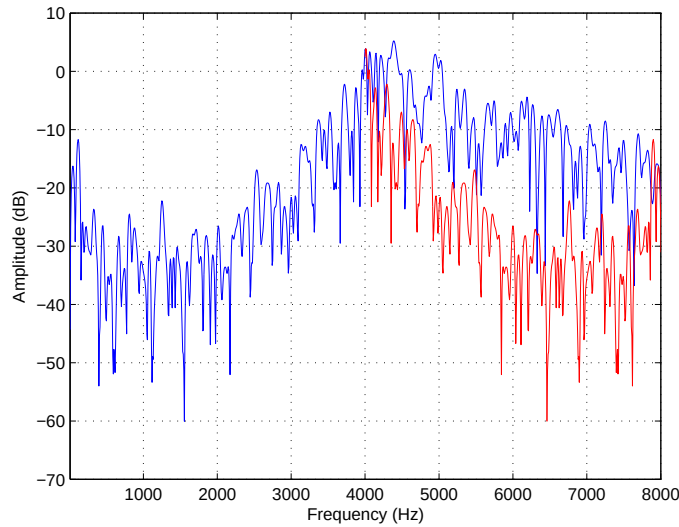


Figure 6.12: Wideband unvoiced excitation signals (blue) and spectrally-mirrored narrowband excitation signal (red) in the frequency domain.

synthesized and added to the narrowband signal to form an estimated wideband signal. Informal listening tests show that this approach achieves a particular improvement in the lowband speech signal. Subjective testing demonstrates that a noticeable improvement in the speech bandwidth is perceived at the expense of introducing some unwanted artefacts.

This discussion is organized as follows. In Section 6.4.1, existing ABWE source methods are reviewed, followed by a description of the proposed data-driven voice source technique. Section 6.4.3 introduces the estimation technique to estimate temporal and spectral envelopes. The system is evaluated in Section 6.4.4.

6.4.1 Existing Voice Source ABWE Techniques

Several methods exist for the artificial bandwidth extension of the high band source signal. Spectral approaches involve translating, mirroring (folding), or modulating the estimated narrowband linear prediction residual, $e^{\text{nb}}(n)$ [113, 186] as shown in Figure 6.12. Techniques that involve filtering and modulating random noise are also employed. Synthetic glottal pulses inserted in synchrony with the long-term predictor in narrowband CODECs can be used in addition to shaped noise [46].

When applied to voiced speech to extend the lowband excitation signal, spectral

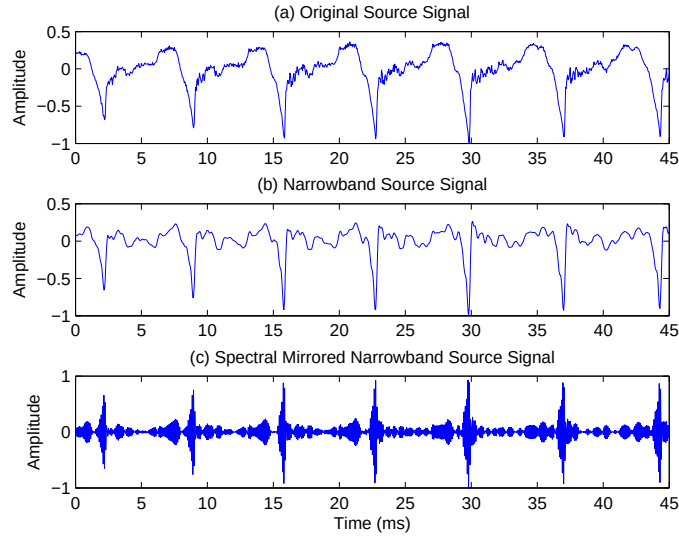


Figure 6.13: Spectral mirroring with voiced speech. (a) Original voice source signal, (b) narrowband voice source signal and (c) spectral mirrored signal.

mirroring techniques generally produce odd time-domain artefacts as shown in Figure 6.13. Existing low band extension techniques include the generation of pitch-synchronous sinusoids [187] or impulse trains [188] and nonlinear processing of $e^{\text{nb}}(n)$ to generate low frequency harmonics with a suitable temporal envelope [189]. Such techniques are generally limited to voiced speech as unvoiced speech contains little energy below 300 Hz. Artefacts associated with lowband extension include buzzing from poorly placed or poorly shaped glottal pulses and roughness caused by incorrectly shaping additive noise.

Existing techniques make little or no use of voice source modelling. The remainder of this section describes an entirely model-based approach based upon DDVSM.

6.4.2 Proposed Model-Based Source Extension

Model Training

Let $s^{\text{wb}}(n) \iff S^{\text{wb}}(z)$ and $s^{\text{nb}}(n) \iff S^{\text{nb}}(z)$ be wideband and narrowband versions of the same speech signal respectively. The corresponding voice sources are $u_D^{\text{wb}}(n)$ and $u_D^{\text{nb}}(n)$. The source signals are divided into amplitude- and scale-normalised frames as

per the method defined in Section 5.2,

$$\begin{aligned}\mathbf{u}_r^{\text{wb}} &= \uparrow_L^K \kappa_r u_D^{\text{wb}}(n'), \\ \mathbf{u}_r^{\text{nb}} &= \uparrow_L^K \kappa_r u_D^{\text{nb}}(n'),\end{aligned}\tag{6.9}$$

where $n' \in \{n_r^c - \frac{1}{2}(n_{r+1}^c - n_{r-1}^c - 2), \dots, n_r^c + \frac{1}{2}(n_{r+1}^c - n_{r-1}^c - 2)\}$, $\uparrow_L^K$ denotes resampling factor $\frac{K}{L}$, L is the number of samples spanned by n' , $K = 2t_{\max}f_s$, $t_{\max}$ corresponds to the maximum glottal period of 0.02 s and κ_r is a gain factor to normalize RMS amplitude. Cycle pairs form the rows of $[R \times K]$ data matrices where R is the total number of cycle pairs,

$$\begin{aligned}\mathbf{U}^{\text{wb}} &= [\mathbf{u}_1^{\text{wb}}, \mathbf{u}_2^{\text{wb}}, \dots, \mathbf{u}_R^{\text{wb}}]^T, \\ \mathbf{U}^{\text{nb}} &= [\mathbf{u}_1^{\text{nb}}, \mathbf{u}_2^{\text{nb}}, \dots, \mathbf{u}_R^{\text{nb}}]^T.\end{aligned}\tag{6.10}$$

An $[R \times F]$ feature matrix of $F = 12$ MFCCs is derived for each narrowband frame,

$$\mathbf{F}^{\text{nb}} = [\mathbf{f}_1^{\text{nb}}, \mathbf{f}_2^{\text{nb}}, \dots, \mathbf{f}_N^{\text{nb}}]^T,\tag{6.11}$$

from which the EM algorithm [133] derives $M = 16$ diagonal covariance Gaussian mixtures. The probability that feature $\mathbf{f}_r^{\text{nb}}$ is a member of mixture component ω_m is stored as an $[R \times M]$ probability matrix with elements $p(\omega_m^{\text{nb}} | \mathbf{f}_r^{\text{nb}})$. For each mixture, the corresponding class centroids, μ_m^{nb} , diagonal covariance matrices, Σ_m^{nb} and mixture weights, $p(\omega_m^{\text{nb}})$, are calculated. The prototype signals are derived as a weighted average of wideband time-domain waveforms, $\mathbf{u}_r^{\text{wb}}$, stored in a $[M \times K]$ matrix,

$$\bar{\mathbf{U}}_{\text{nb}}^{\text{wb}} = [\bar{\mathbf{u}}_{\text{nb},1}^{\text{wb}}, \bar{\mathbf{u}}_{\text{nb},2}^{\text{wb}}, \dots, \bar{\mathbf{u}}_{\text{nb},M}^{\text{wb}}]^T = \mathbf{P}_{\text{nb}}^T \mathbf{U}^{\text{wb}}.\tag{6.12}$$

We employ a convention for $\mathbf{U}$ and $\mathbf{P}$ whereby the superscript refers to the bandwidth of the time-domain waveforms and the subscript to that of the feature set. The system is depicted in Figure 6.14.

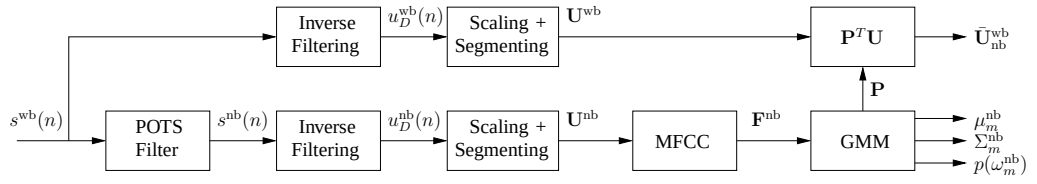


Figure 6.14: System diagram for training of the proposed excitation source signal estimator.

Wideband Voice Source Estimation

Wideband voice source estimation is similar to model training. A narrowband test utterance is inverse-filtered and segmented into amplitude and scale-normalized 2-cycle frames, $\mathbf{u}_r^{\text{nb}}$, with corresponding MFCCs, $\mathbf{f}_r^{\text{nb}}$. The DYPSA algorithm [77] provides estimation of GCIs for segmentation. The decomposition for frame r is

$$\boldsymbol{\varphi}_r = [\varphi_{1,r}, \varphi_{2,r}, \dots, \varphi_{M,r}]^T = [p(\omega_1^{\text{nb}} | \mathbf{f}_r^{\text{nb}}), p(\omega_2^{\text{nb}} | \mathbf{f}_r^{\text{nb}}), \dots, p(\omega_M^{\text{nb}} | \mathbf{u}_r^{\text{nb}})]^T. \quad (6.13)$$

We define a set $\Gamma_r \subseteq \Gamma_{\text{all}}$, $\Gamma_{\text{all}} = \{1, \dots, M\}$. Γ_r contains the class indices that produce the highest likelihood. A wideband cycle of the voice source signal can then be resynthesized from the prototypes, $\bar{\mathbf{U}}_{\text{nb}}^{\text{wb}}$, with the decomposition terms,

$$\hat{\mathbf{u}}_r^{\text{wb}} = \sum_{m \in \Gamma_r} \varphi_{m,r} \left(\uparrow_{\alpha}^{\beta} \kappa_r \bar{\mathbf{u}}_{\text{nb},m}^{\text{wb}} \right), \quad (6.14)$$

where $\uparrow_{\alpha}^{\beta}$ resamples $\bar{\mathbf{u}}_m$ to length L and κ_r is a gain factor to reproduce the same energy as the source cycle. An approximation to the full $u_D(n)$ is synthesized by windowing, shifting and summing as described in Section 5.2.1.

The proposed approach is suitable for voiced speech only; unvoiced excitation is produced with spectral mirroring. An untrained voiced/unvoiced/silence detector is employed as described in Section 6.2.2.

6.4.3 Proposed Temporal & Spectral Envelope Estimation

To complete the ABWE scheme, the wideband signal envelope has to be estimated and restored. Several envelope parameterizations have been proposed for ABWE. Mostly, an

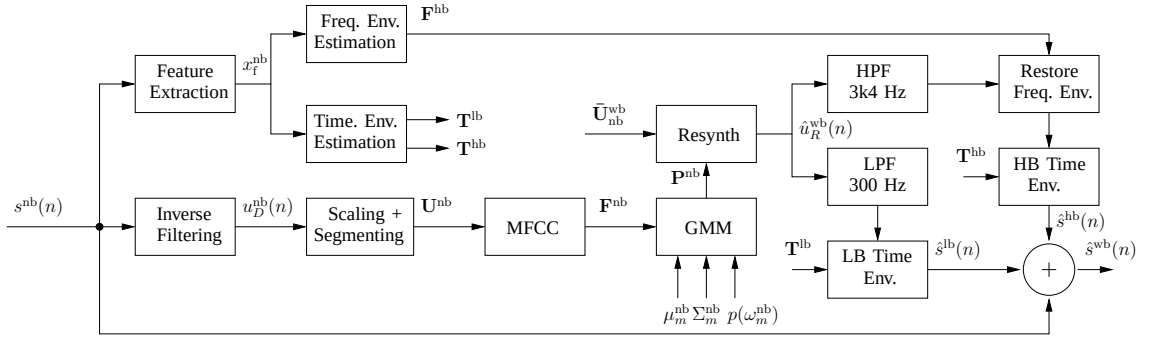


Figure 6.15: System diagram for ABWE synthesis.

autoregressive model is assumed and the envelope is restored using an LPC synthesis filter with estimated coefficients. However, LPC synthesis of artificial source signals does not necessarily regenerate the correct temporal characteristics. Therefore, in this work, a signal parameterization is employed in terms of spectral *and* temporal energy envelopes [46], whereby low and high extension bands are treated separately. For the high extension band, the spectral envelope is parameterized in terms of 10 logarithmic subband energies, F^{hb} , (375 Hz subbands) for each 10 ms frame. The temporal envelope, T^{hb} , provides 5 logarithmic subframe energies of the extension band signal for each 10 ms frame (2 ms subframes). For the low extension band, only the temporal envelope, T^{lb} , of the low-pass signal is used.

The parameter vectors F^{hb} , T^{hb} and T^{lb} are estimated with separate HMM-based MMSE estimators [113]. The estimators require a narrowband feature vector, x_f^{nb} , for each frame. Here, x_f^{nb} is composed of the narrowband MFCCs, of the zero crossing rate and of the narrowband temporal envelope T^{nb} . The actual estimator configurations are listed in Table 6.2. Based on the estimated parameter set, the extension band signals $\hat{s}^{hb}(n)$ and $\hat{s}^{lb}(n)$ can be synthesised by shaping the envelopes of the source signals, $\hat{u}^{hb}(n)$ and $\hat{u}^{lb}(n)$, respectively. This signal shaping is performed in a two-step approach: a filterbank equalizer restores the spectral envelope (high band only) and the temporal envelope is corrected via gain manipulation, cf. [46]. Finally, the signals $\hat{s}^{hb}(n)$ and $\hat{s}^{lb}(n)$ are combined with $s^{nb}(n)$ to give the bandwidth extended output $\hat{s}^{wb}(n)$. The estimation / resynthesis procedure is shown in Figure. 6.15.

6.4.4 Evaluation

Four voice source estimation techniques were considered for subjective testing: i) spectral mirroring, ii) synthetic glottal pulse located at the GCIs during voiced + spectral mirroring during unvoiced, iii) DDVSM during voiced + spectral mirroring during unvoiced and iv) DDVSM for LB + spectral mirroring for HB. The ABWE techniques were applied to narrowband speech, quantized with an ITU-T G.711 μ -law audio CODEC [190].

An ITU-T P.800 [117] subjective test was devised, including two additional hidden references in the form of wideband and quantized narrowband speech. The sample set was a subset of the NTT database [121], consisting of 3 female and 3 male talkers, each speaking 5 pairs of phonetically-balanced sentences. The 20 subjects each listened to the 30 samples in random order, with one of the 6 methods randomly applied to each sentence. Processed sentences were normalized to a level of -30 dB with respect to the overload point defined in ITU-T P.56 [118], then presented with Sennheiser HD650 headphones in a listening room environment. Subjects were asked to rate i) *Foreground*, describing speech quality only, ii) *Background*, describing artefact tolerance, and iii) *Overall* impression. An Absolute Category Rating (ACR) scale was used for i) and iii) and a Degradation Category Rating (DCR) for ii), rated 1 – 5 in 0.5 increments. Five examples were given with approximate ratings prior to taking the test. A set of control samples, rated by a team of expert listeners, were used to derive a quadratic calibration curve for each subject to standardize their responses.

Table 6.2: Wideband Envelope Estimator Configurations

Param. Vect.	Param. Dim.	# of Features	Codebook Size	# of Gaussians.
$\mathbf{F}_{\text{hb}}$	10	19	128	8
$\mathbf{T}_{\text{hb}}$	5	19	128	8
$\mathbf{T}_{\text{lb}}$	5	19	64	8

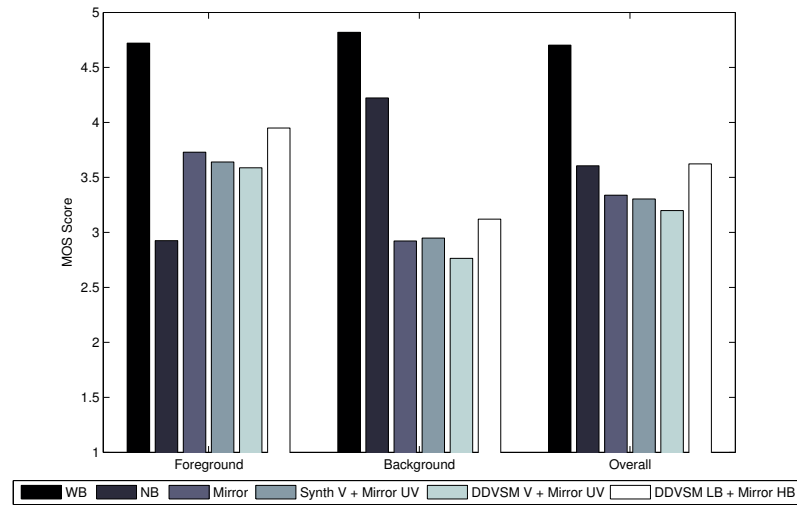


Figure 6.16: MOS results for ABWE algorithms.

Results and Discussion

The results show that all ABWE techniques improve the perceived foreground score at the expense of reducing the background score. Of the techniques under test, a clear preference was shown for the combined DDVSM LB + spectral mirroring HB, confirming the assertion that DDVSM is particularly effective for lowband extension and that lowband ABWE is especially sensitive to the source signal employed. The preference of the best ABWE technique compared with narrowband is still relatively small. These results contrast with previous findings where highband-only ABWE is preferred to narrowband, suggesting that lowband artefacts are particularly detrimental to perceived quality.

Two artefacts regularly occurred in the test set. The first was a *beating* in the low extension band, caused by erroneous GCI detections, resulting in a low frequency excitation signal that was not pitch-synchronous with the narrowband signal. The second was a *hissing* in the high extension band, caused by incorrect estimation of the upper spectral envelope. Such artefacts are expected to be reduced by fine-tuning of the training and estimation process. Examples of ABWE speech can be found at [183].

6.5 Chapter Summary

Three applications to existing problems in speech processing using glottal-synchronous techniques have been discussed in this chapter.

The Spatiotemporal Method for Enhancement of Reverberant Speech (SMERSH) is a method for speech dereverberation that employs spatial filtering with multi-microphone observations and temporal filtering by averaging neighbouring glottal cycles. The Multichannel DYPSA (MC-DYPSA) algorithm, first discussed in Section 4.5.3, is used for the accurate estimation of GCIs from multichannel reverberant speech recordings. Clean speech samples, played through a speaker and recorded in a reverberant office environment, show that improvements in segmental Segmental Signal-to-Noise Ratio (SSNR) of up to 5 dB and 0.25 BSD can be achieved.

Time scale modification is the altering of the length of a speech segment without changing pitch, prosody or formant structure. A method for time scale modification has been proposed that gives good perceptual quality for a range of modification factors as demonstrated by subjective testing. The method employs an unvoiced / transition (UT) detector which ensures that time scale modification is only applied to silence or voiced segments, avoiding the artefacts caused by the time scale modification of unvoiced and transition sounds. GCIs are provided by the DYPSA algorithm as part of a practical, fast, reliable approach for time scale modification. The method was tested by subjective MOS testing against the same approach but excluding the UT detector. The results suggest that UT detection is preferred, though more so at larger modification factors. At small modification factors the benefit of UT detection is less pronounced. The ability to alter the rate of speech reliably can enable applications such as time compression for the fast scanning of recorded messages, expansion for improving intelligibility or a mixture such as in the case of audio-video synchronization.

Artificial Bandwidth Extension (ABWE) is the process of estimating missing frequency components from narrowband speech recordings. An ABWE technique has been proposed that employs spectral mirroring and Data-Driven Voice Source Modelling

(DDVSM) to estimate a wideband source signal from narrowband speech. Used in conjunction with a state-of-the-art framework that estimates the temporal and spectral envelopes of the extension bands, the ABWE system is novel in its explicit use of voice source modelling and the estimation both the low ($50 - 300$ Hz) the high ($3.4 - 7$ kHz) extension bands. Informal listening tests reveal that the proposed technique is particularly effective in the lowband. Formal subjective tests demonstrate that an improvement in the perceived bandwidth of speech can be achieved at the expense of increasing background artefacts. It further reveals that, compared with the other methods under test, the use of DDVSM in the lowband with spectral mirroring in the highband is preferred over narrowband speech, as it provides the greatest perceived ABWE with the least number of unwanted artefacts.

Chapter 7

Conclusions

THIS chapter summarises and concludes the work presented in this thesis. Section 7.1 contains a résumé of the main problems addressed with some important results. Further work is proposed in Section 7.2. Publications arising from this thesis are detailed in Section 7.3.

7.1 Résumé

Glottal-synchronous speech processing is a field of speech science that exploits the pseudoperiodicity of voiced speech. Central to this field of research is the source-filter model which describes speech production as a linear combination of a pseudoperiodic glottal excitation signal (the voice source) that excites a vocal tract filter. Methods for parameterising and inverting the vocal tract filter are well-established, although the nature of the voice source is less well-understood owing to the relatively challenging problem of defining effective models. This work aimed to demonstrate that through the accurate detection of periodicity in the voice source, often delimited by the glottal closure instants, improved models of the voice source may be established and some long-standing challenges in speech signal processing addressed. The key contributions of this work are outlined as follows:

Detection of Glottal Activity from EGG Signals Glottal Closure Instants (GCIs), and to a lesser extent the Glottal Opening Instants (GOIs), are the foundation for

glottal-synchronous speech processing. Detection of these instants from speech signals is a challenging task due to spectral filtering by the vocal tract, additive acoustic noise from nearby sources and/or recording equipment and convolutive noise from acoustic reverberation. Such distortions may be circumvented by invasive measures such as the Electroglottograph (EGG), or Laryngograph. A few algorithms exist that achieve reliable detection during voiced speech segments but many fail at the transitions from unvoiced to voiced and voiced to unvoiced. The SIGMA algorithm was proposed that is reliable for natural conversational speech, giving GCI and GOI detection rates of 99.59% and 99.47% against a hand-labelled reference respectively, making it suitable for use as a reference for detection from speech signals and for use in certain areas of pathological speech.

Detection of Glottal Activity from Speech Signals In many real-world situations, invasive measurements are not available so GCIs and GOIs must be detected from speech directly. A number of speech-based GCI detection algorithms exist but few are suitable for natural conversational speech and fewer still are capable of extracting GOIs. The YAGA algorithm was proposed for the reliable detection of GCIs and GOIs from clean and noisy speech. Identification rates of 99.84% on clean speech were recorded against a hand-labelled reference, with similar results for SNR > 10 dB. Detection from reverberant environments is more challenging as it produces many more incorrect candidates. A multichannel extension to the DYPSA algorithm was proposed that exploits the spatial diversity of reverberant speech to produce a cost function based upon interchannel correlation of candidates. Those candidates that are highly-correlated across channels are shown to be more likely to correspond to true GCIs than those that are not. Evaluation results show that with a 0.5 s reverberation time, the identification rate of the proposed algorithm exceeds that of DYPSA by 18%.

Data-Driven Voice Source Modelling Models of the voice source signal are often based upon fitting polynomial, exponential or sinusoidal curves to observations, physical models of coupled damped mass-spring systems or empirical derivation by

minimising an error measure between observed and signals synthesised with noise and/or glottal pulse codebooks. Many are unable to produce the whole gamut of sounds produced by a real voice source. The use of training data to derive models is limited owing to the relative difficulty of reliably segmenting a voice source signal into individual glottal cycles; the techniques developed in this work enable such segmentation. By deriving a data matrix of voice source cycles that have been normalised in scale and amplitude to remove dependence upon speaker pitch and volume, machine learning techniques can be applied to extract long-term trends in the voice source. A compact description that is capable of reliably reproducing a voice source signal was demonstrated. Examples of further applications in LPC pre-emphasis, analysis, speaker identification, compression and enhancement were given to demonstrate the wide applicability of the approach. The compression experiment showed that a reconstruction SSNR of 12 dB is achieved with $1/8$ the full spectra, with perceptually perfect reconstruction for $>1/3$ the full spectra.

Dereverberation Providing GCIs are reliably detected, a number of long-standing speech processing problems can be addressed. The Spatiotemporal Method for Enhancement of Reverberant Speech (SMERSH) algorithm is an existing technique for enhancing reverberant speech by exploiting the spatial diversity and pseudoperiodicity of speech captured with multiple microphones. The Multichannel DYPSA Algorithm, proposed in this work, was shown to enable the practical application of SMERSH to real-world signals, producing enhanced speech with both objective and subjective measures, exceeding that produced by existing spatial averaging techniques. Improvements in SSNR of up to 5 dB and 0.25 BSD over the reverberant speech were recorded.

Time-scale modification The lengthening or shortening of a speech signal without affecting its formant structure can be achieved with the Pitch-Synchronous Overlap-Add (PSOLA) algorithm, where individual cycles are repeated or removed depending whether an expansion or compression of time scale is desired. However, the processing of unvoiced speech in this way introduces annoying artefacts. A method

was proposed that applied the DYPSA algorithm to estimate GCIs and a trained voiced/unvoiced/silence detector ensured that unvoiced speech was left unprocessed. A significant perceptual improvement over standard PSOLA was measured with formal subjective tests.

Artificial Bandwidth Extension Telephone speech is bandwidth-limited to 300 Hz – 3.4 kHz, significantly impairing intelligibility compared with wideband speech (50 Hz – 7 kHz). Artificial Bandwidth Extension (ABWE) is a technique for replacing the missing frequency components to improve the quality of narrowband speech, for which existing techniques for estimating the upper extension band have been successful. It can be demonstrated that in the upper extension band the spectral envelope is of greater perceptual significance than source signal; the converse applies in the low extension band for which the source signal is of greater importance. Glottal-synchronous techniques developed in this work apply a mapping from narrowband features to wideband source signals, which are capable of producing perceptually superior results when used to estimate the lowband source signal as demonstrated by subjective tests.

7.2 Further Work

The development of the YAGA and SIGMA algorithms for GCI detection were the result of many incremental refinements, engendered through flaws revealed with real-world applications on natural conversational speech. With reliable and accurate GCI detection algorithms in place, it is hoped that future work in the field of glottal-synchronous speech processing can have a greater focus on practical applications.

Voice Source Estimation Techniques for estimating the voice source are all, to some degree, unable to completely dissociate the voice source from the vocal tract filter due to the effects of nonlinearity and non-whiteness of the voice source. The use of an LPC preemphasis filter in the form of the least-squares inverse of average glottal pulse was shown to greatly improve the accuracy of a GCI detector. A method for

choosing a dynamic preemphasis filter based upon trained models of voiced speech may improve results further. Investigation of voice source estimation with closed-phase analysis using GCIs and GOIs estimated with YAGA is an area for further research.

DDVSM Optimization Criteria The key to DDVSM is the construction of a data matrix containing voice source cycles that are normalised in scale and amplitude. The methods presented in this thesis are largely ‘off-the-shelf’ algorithms but they have demonstrated that the technique can be applied effectively to many areas of speech science. Investigation into the optimization criteria for the learning algorithms may yield perceptually superior results.

Coding An experiment with voice source reconstruction from a truncated set of PCA spectra showed that the technique can yield a compact and accurate representation of the voice source signal. A second experiment with the clustering of PCA spectra showed that the voice source for natural conversational speech is largely piecewise-constant. A combination of the two results form the basis of an efficient coding scheme. Further investigation should be made into its suitability for a high-quality low-bitrate coder or a toll-quality ultra low-bitrate coder.

Speaker Identification An experiment with clustering of PCA spectra of the voice source revealed that certain classes can be an estimator of speaker. Further experimentation with a wider training set and alternative decomposition methods will help to ascertain whether speaker identification can be reliably performed using voice source signals.

7.3 Publications

The following publications were produced during the course of this work:

7.3.1 Conferences

1. M. R. P. Thomas, N. D. Gaubitch and P. A. Naylor, “Multichannel DYPSA for estimation of glottal closure instants in reverberant speech,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Poznan, Poland, Sept. 2007.
2. M. R. P. Thomas, N. D. Gaubitch, J. Gudnason and P. A. Naylor, “A practical multichannel dereverberation algorithm using Multichannel DYPSA and Spatiotemporal Averaging,” in *Proc. IEEE Workshop on App. of Signal. Processing to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2007.
3. N. D. Gaubitch, M. R. P. Thomas and P. A. Naylor, “Subband Method for Multichannel Least Squares Equalization of Room Transfer Functions,” in *Proc. IEEE Workshop on App. of Signal. Processing to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2007.
4. M. R. P. Thomas and P. A. Naylor, “The SIGMA Algorithm for Estimation of Reference-Quality Glottal Closure Instants from Electroglottograph Signals,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, August 2008.
5. M. R. P. Thomas, J. Gudnason and P. A. Naylor, “Application of the DYPSA Algorithm to Segmented Time Scale Modification of Speech,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, August 2008.
6. M. R. P. Thomas, J. Gudason and P. A. Naylor, “Data-Driven Voice Source Waveform Modelling,” in *Proc. International Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.
7. M. R. P. Thomas, J. Gudnason and P. A. Naylor, “Detection of Glottal Closing

and Opening Instants Using the Improved DYPSA Framework,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Scotland, August 2009.

8. J. Gudnason, M. R. P. Thomas, P. A. Naylor and D. P. W. Ellis, “Voice Source Waveform Analysis and Synthesis using Principal Component Analysis and Gaussian Mixture Modelling,” in *Proc. Interspeech Conf*, Brighton, UK, Sept. 2009.
9. M. R. P. Thomas, B. Geiser, J. Gudnason, P. A. Naylor and P. Vary, “Voice Source Estimation for Artificial Bandwidth Extension of Telephone Speech,” to appear in *Proc. International Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, USA, March 2010.

7.3.2 Journals

1. M. R. P. Thomas and P. A. Naylor, “The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8 pp. 1557-1566, Nov. 2009.
2. M. R. P. Thomas, J. Gudnason and P. A. Naylor, “Detection of Glottal Closing and Opening Instants With the YAGA Algorithm,” *Submitted to IEEE Trans. Audio, Speech, Lang. Process.*.
3. J. Gudnason, M. R. P. Thomas, P. A. Naylor and D. P. W. Ellis, “On Data-Driven Voice Source Modelling,” In Progress.

7.3.3 Contributions to Books

1. N. D. Gaubitch, M. R. P. Thomas and P. A. Naylor, “Dereverberation using LPC-Based Approaches,” to appear in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, (eds.), Springer, 2010, ch. 4, pp. 99–132.

Bibliography

- [1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
- [2] H. W. Strube, “Determination of the instant of glottal closure from the speech wave,” *J. Acoust. Soc. Am.*, vol. 56, no. 5, pp. 1625–1629, 1974.
- [3] R. Colton and J. Casper, *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. New York: Williams & Wilkins, 1996.
- [4] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Frances, 2000.
- [5] H. Pulakka, P. Alku, L. Laaksonen, and P. Valve, “The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech,” in *Proc. Interspeech Conf.*, Antwerp, Belgium, Aug. 2007, pp. 2497–2500.
- [6] O’Shaughnessy, *Speech Communication: Human and Machine*. Addison Wesley, 1987.
- [7] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, “Laryngographic assessment of normal voice: a tutorial,” *Clinical Linguistics and Phonetics*, vol. 3, pp. 281–296, 1989.
- [8] P. Davies, G. A. Lindsey, H. Fuller, and A. J. Fourcin, “Variation of glottal open and closed phases for speakers of English,” *Proc. Institute of Acoustics*, vol. 8, no. 7, pp. 539–546, 1986.
- [9] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*. Prentice Hall, 1993.

- [10] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.
- [11] V. N. Sorokin, "Determination of vocal tract shape for vowels," *Speech Communication*, vol. 11, pp. 71–85, 1992.
- [12] T. V. Ananthapadmanabha and G. Fant, "Calculations of true glottal volume-velocity and its components," *Speech Communication*, vol. 1, pp. 167–184, 1982.
- [13] M. M. Sondhi, "Model for wave propagation in a lossy vocal tract," *J. Acoust. Soc. Am.*, vol. 55, pp. 1070–1075, May 1974.
- [14] J. Thyssen, H. Nielsen, and S. D. Hansen, "Non-linear short-term prediction in speech coding," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 1994, pp. 185–188.
- [15] P. Maragos, T. F. Quatieri, and J. F. Kaiser, "Speech nonlinearities, modulations, and energy operators," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 1991, pp. 421–424.
- [16] H. K. Dunn, "The calculation of vowel resonances, and an electrical vocal tract," *J. Acoust. Soc. Am.*, vol. 22, no. 12, pp. 740–753, 1950.
- [17] P. M. Morse, *Vibration and Sound*. New York: McGraw Hill, 1948.
- [18] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey, USA: Prentice-Hall, 1978.
- [19] G. Fant, "Analysis and synthesis of speech processes," in *Manual of Phonetics*, B. Malmberg, Ed. North Holland, 1968, ch. 8, pp. 173–276.
- [20] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. of the Fourth Intl. Congress on Acoustics*, vol. G42, 1962, pp. 1–4.
- [21] L. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1988.

- [22] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamlin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of cs-acelp: A toll quality 8 kb/s speech coder," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 116–129, Mar. 1998.
- [23] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637–655, Aug. 1971.
- [24] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Springer-Verlag, 1976.
- [25] O. O. Akande and P. J. Murphy, "Estimation of the vocal tract transfer function with application to glottal wave analysis," *Speech Communication*, vol. 46, no. 1, pp. 15–36, May 2005.
- [26] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.
- [27] H. Strik, B. Cranen, and L. Boves, "Fitting LF-model to inverse filtered signals," in *Eurospeech*, vol. 1, Berlin, 1993, pp. 103–106. [Online]. Available: <http://lands.let.kun.nl/TSPublic/strik/publications>
- [28] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.*, vol. 49, pp. 583–590, Feb. 1971.
- [29] G. Fant and J. Liljencrants, "Perception of vowels with truncated intraperiod decay envelopes," in *STL-QPSR 1*. <http://www.speech.kth.se>: Department of Speech, Music and Hearing, KTH, 1979, pp. 79–84.
- [30] G. Fant, "Vocal source analysis, a progress report," in *STL-QPSR 3–4*. <http://www.speech.kth.se>: Department of Speech, Music and Hearing, KTH, 1979, pp. 31–53.
- [31] —, "Glottal flow: models and interaction," *J. Phonet*, vol. 14, pp. 393–399, 1986.

- [32] D. M. Brookes and D. S. Chan, "Speaker characteristics from a glottal airflow model using glottal inverse filtering," *Proc. Institute of Acoustics*, vol. 15, pp. 501–508, 1994.
- [33] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–576, Sept. 1999.
- [34] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, Feb. 1990.
- [35] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 12, 1987, pp. 637–640.
- [36] P. Alku and T. Backstrom, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.
- [37] J. Flanagan and L. Landgraf, "Self-oscillating source for vocal-tract synthesizers," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 57–64, Mar. 1968.
- [38] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*. Springer-Verlag New York, 1991.
- [39] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.
- [40] D. E. Dudgeon, "Two-mass model of the vocal cords," *J. Acoust. Soc. Am.*, vol. 48, p. 118A, 1970.
- [41] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, pp. 1249–1260, 1994.

- [42] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Paris, France, May 1982, pp. 614–617.
- [43] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 937–940.
- [44] A. Bergstrom and P. Hedelin, "Code-book driven glottal pulse analysis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1989, pp. 53–56.
- [45] D. McElroy, B. P. Murray, and A. D. Fagan, "Wideband speech coding using multiple codebooks and glottal pulses," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 253–256.
- [46] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaume, and S. Ragot, "Bandwidth extension for hierarchical speech and audio coding in ITU-T rec. G.729.1," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.
- [47] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Am.*, vol. 31, no. 6, pp. 667–677, June 1959.
- [48] J. D. Markel, "Application of a digital inverse filter for automatic formant and f_0 analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 3, pp. 154–160, June 1973.
- [49] J. Lindqvist-Gauffin, "Studies of the voice source by means of inverse filtering," *STL-QPSR*, vol. 6, no. 2, pp. 8–13, 1965.
- [50] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
- [51] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

- [52] K. Schnell and A. Lacroix, "Weighted nonlinear prediction based on volterra series for speech analysis," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.
- [53] P. Chytil and M. Pavel, "Variability of glottal pulse estimation using cepstral method," in *Proc. 7th Nordic Signal Processing Symposium (NORSIG)*, 2006, pp. 314–317.
- [54] D. Y. Wong, J. D. Markel, and J. A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.
- [55] A. H. Gray and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 22, pp. 207–217, 1974.
- [56] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 369–377, Apr. 1985.
- [57] D. S. F. Chan and D. M. Brookes, "Variability of excitation parameters derived from robust closed phase glottal inverse filtering," *Proc. European Conf. on Speech Communication and Technology*, vol. 33, no. 1, Sept. 1989.
- [58] M. Rothenberg, "A new inverse filtering technique for deriving the glottal airflow waveform during voicing," *J. Acoust. Soc. Am.*, vol. 53, pp. 1632–1645, 1973.
- [59] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 28–42, Feb. 1986.
- [60] A. P. Lobo, "Glottal flow derivative modeling with the wavelet smoothed excitation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 2001, pp. 861–864.

- [61] D. G. Childers, J. C. Principe, and Y. T. Ting, "Adaptive WRLS-VFF for speech analysis," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 209–213, May 1995.
- [62] W. Ding and H. Kasuya, "A novel approach to the estimation of voice source and vocal tract parameters from speech signals," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, vol. 2, 1996, pp. 1257–1260.
- [63] W. Zhu and H. Kasuya, "A new speech synthesis system based on the ARX speech production model," in *Proc. Intl. Conf. on Spoken Language Processing*, 1996, pp. 1413–1416. [Online]. Available: <http://www.asel.udel.edu/icslp/cdrom/icslp96.htm>
- [64] D. G. Childers, D. P. Skinner, and R. C. Kemerait, "The cepstrum: A guide to processing," *Proc. IEEE*, vol. 65, no. 10, pp. 1428–1443, 1977.
- [65] A. Petropulu and S. Subramaniam, "Cepstrum based deconvolution for speech dereverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1994, pp. I–12.
- [66] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 392–396, 1996.
- [67] R. Kennedy and B. Radlović, "Iterative cepstrum-based approach for speech dereverberation," in *Proc. fifth Intl. Symposium on Signal Processing and its Applications (ISSPA)*, vol. 1, 1999, pp. 55–58.
- [68] D. G. Childers, D. M. Hooks, G. P. Moore, L. Eskenazi, and A. L. Lalwani, "Electroglottography and vocal fold physiology," *J. Speech. Hear. Res.*, vol. 33, no. 2, pp. 245–254, June 1990.
- [69] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1557–1566, Nov. 2009.
- [70] J. Van den Berg, "Modern research in experimental phoniatrics," *Folia Phoniatrica*, vol. 14, pp. 81–149, 1962.

- [71] M. Rothenberg, "Measurement of airflow in speech," *Journal of Speech, Language and Hearing Research*, vol. 20, no. 1, pp. 155–176, Mar. 1977.
- [72] P. Badin, S. Hertegard, and I. Karlsson, "Notes on the Rothenberg Mask," *STL-QPSR*, vol. 31, no. 1, pp. 1–7, 1990.
- [73] D. Bless, M. Hirano, and R. Feder, "Videostroboscopic evaluation of the larynx," *Ear, Nose and Throat Journal*, vol. 66, pp. 48–58, July 1987.
- [74] S. Kiritani, H. Imagawa, and H. Hirose, "High-speed digital image recording for the observation of vocal cord vibration," in *Voice Physiology: Voice Production, Mechanism, and Functions*, O. Fujimura, Ed. New York: Raven Press, 1988, pp. 261–269.
- [75] J. Švec and H. Schutte, "Videokymography: High-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, pp. 201–205, 1996.
- [76] C. Drioli and F. Avanzini, "Hybrid parameteric-physiological glottal modelling with applicatino to voice quality assessment," *Journal of Medical Engineering & Physics*, vol. 24, no. 7–8, pp. 453–460, Sept. 2002.
- [77] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.
- [78] M. Brookes, P. A. Naylor, and J. Gudnason, "A quantitative assessment of group delay methods for identifying glottal closures in voiced speech," *IEEE Trans. Speech Audio Process.*, vol. 14, 2006.
- [79] R. Schwartz and Y.-L. Chow, "The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1990, pp. 81–84.
- [80] R. C. Scherer, V. J. Vail, and B. Rockwell, "Examination of the laryngeal adduction measure EGGW," in *Producing Speech: Contemporary Issues: for Katherine Safford*

- Harris*, F. Bell-Berti and L. J. Raphael, Eds. American Institute of Physics, 1995, pp. 269–290.
- [81] B. Atal and L. Rabiner, “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 3, pp. 201–212, June 1976.
- [82] H. Maqsood and P. A. Naylor, “Improved DYPSA algorithm for noise and unvoiced speech,” in *Proc. Intl. Conf. on Emerging Technologies*, Islamabad, Nov. 2007, pp. 243–248.
- [83] S. Furui, “Comparison of speaker recognition methods using statistical features and dynamic features,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, pp. 342–350, June 1981.
- [84] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, “Application of the DYPSA algorithm to segmented time-scale modification of speech,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.
- [85] C. Hoelper, A. Frankort, C. Erdmann, and P. Vary, “A novel voiced / unvoiced / silence classification scheme for offline speech coding,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, vol. 3, Toulouse, France, Sept. 2002, pp. 121–124.
- [86] A. de Cheveigne and H. Kawahara, “Comparative evaluation of F0 estimation algorithms,” in *Proc. Eurospeech*, 2001.
- [87] L. Malfait, J. Berger, and M. Kastner, “P.563 - the ITU-T standard for single-ended speech quality assessment,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, 2006.
- [88] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn, “Low-complexity, non-intrusive speech quality assessment,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1948–1956, Nov. 2006.

- [89] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995, pp. 495–518.
- [90] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.
- [91] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.
- [92] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, vol. 21, pp. 577–580, 1949.
- [93] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.
- [94] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.
- [95] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.
- [96] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 541–544.
- [97] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

- [98] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 3, pp. 374–390, June 1981.
- [99] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, Dec. 1990.
- [100] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175–205, Feb. 1995.
- [101] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.
- [102] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1993, pp. 554–557.
- [103] J. di Martino and Y. Laprie, "Suppression of phasiness for time-scale modifications of speech signals based on a shape invariance property," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 2001, pp. 853–856.
- [104] T. Ebihara, Y. Ishikawa, Y. Kisuki, T. Sakamoto, and T. Hase, "Speech synthesis software with variable speaking rate and its implementation on a 32-bit microprocessor," in *Proc. Intl. Conf. on Consumer Electronics ICCE*, June 2000, pp. 254–255.
- [105] M. Demol, W. Verhelst, K. Struyve, and P. Verhoeve, "Efficient non-uniform time-scaling of speech with WSOLA," in *Proc. Speech and Computers*, Patras, Greece, Oct. 2005, pp. 163–166.

- [106] O. Donnellan, E. Jung, and E. Coyle, "Speech-adaptive time-scale modification for computer assisted language-learning," in *Proc. IEEE Intl. Conf. on Advanced Learning Technologies*, July 2003, pp. 165–169.
- [107] S. Lee, H. D. Kim, and H. S. Kim, "Variable time-scale modification of speech using transient information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 1997, pp. 1319–1322.
- [108] D. Kapilow, Y. Stylianou, and J. Schroeter, "Detection of non-stationarity in speech signals and its application to time-scaling," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999, pp. 2307–2310.
- [109] M. Covell, M. Withgott, and M. Slaney, "MACH1: nonuniform time-scale modification of speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1998, pp. 349–352.
- [110] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Edinburgh, Scotland, Sept. 1994, pp. 1178–1181.
- [111] Y. Nakatoh, M. Tushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. European Conf. on Speech Communication and Technology*, vol. 3, Rhodes, Greece, Sept. 1997, pp. 1643–1646.
- [112] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformations," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, June 2000, pp. 1843–1846.
- [113] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, 2003.
- [114] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.

- [115] A. S. Spanias, "Speech coding: a tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.
- [116] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing For Telecommunications*, S. L. Gay and J. Benesty, Eds. Kluwer Academic Publishers, 2000.
- [117] *Methods for subjective determination of transmission quality*, Online, International Telecommunications Union (ITU-T) Recommendation P.800, Aug. 1996. [Online]. Available: <http://www.itu.int/rec/T-REC-P.800/en>
- [118] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.
- [119] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London, Technical Report, June 1987.
- [120] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *Proc. European Conf. on Speech Communication and Technology*, Sept. 1995, pp. 867–870.
- [121] N. A. T. Corporation, "Multi-lingual speech database for telephonometry 1994." [Online]. Available: http://www.ntt-at.com/products_e/speech/
- [122] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Multichannel DYPSA for estimation of glottal closure instants in reverberant speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Poznan, Poland, Sept. 2007.
- [123] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.

- [124] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
- [125] W. Saidi, A. Bouzid, and N. Ellouze, "Evaluation of multi-scale product method and DYPSA algorithm for glottal closure instant detection," in *Proc. 3rd Intl. Conf. on Information and Communication Technologies: From Theory to Applications (ICTTA)*, Apr. 2008, pp. 1–5.
- [126] K. Verdolini, R. Chan, I. R. Titze, M. Hess, and W. Bierhals, "Correspondence of electroglottographic closed quotient to vocal fold impact stress in excised canine larynges," *Journal of Voice*, vol. 12, no. 4, pp. 415–423, Feb. 1998.
- [127] J. Gamboa, F. J. Jiménez-Jiménez, A. Nieto, I. Cobeta, A. Vegas, M. Orti-Pareka, T. Gasalla, J. A. Molina, and E. Garcia-Albea, "Acoustic voice analysis in patients with essential tremor," *Journal of Voice*, vol. 12, no. 4, pp. 444–452, Feb. 1998.
- [128] D. M. Howard, "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers," *Journal of Voice*, vol. 9, no. 2, pp. 121–1223, June 1995.
- [129] N. Henrich, C. d'Alessandro, M. Castellengo, and B. Doval, "On the use of the derivative of electroglottographic signals for characterization of nonpathological voice phonation," *J. Acoust. Soc. Am.*, vol. 115, no. 3, pp. 1321–1332, Mar. 2004.
- [130] M. A. Huckvale, "Speech filing system: Tools for speech," University College London, Tech. Rep., 2004. [Online]. Available: <http://www.phon.ucl.ac.uk/resource/sfs>
- [131] A. Bouzid and N. Ellouze, "Multiscale product of electroglottogram signal for glottal closure and opening instant detection," in *Proc. IMACS MultiConf. on Computational Engineering in Systems Applications*, vol. 1, 2006, pp. 106–109.
- [132] —, "Glottal opening instant detection from speech signal," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Sept. 2004, pp. 729–732.

- [133] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [134] M. B. Higgins and J. H. Saxman, "A comparison of selected phonatory behaviours of healthy aged and young adults," *Journal of Speech and Hearing Research*, vol. 34, pp. 1000–1010, Oct. 1991.
- [135] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *J. Speech. Hear. Res.*, vol. 31, no. 3, pp. 338–351, Sept. 1988.
- [136] A. Bouzid and N. Ellouze, "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," *Speech Communication*, vol. 51, pp. 782–792, Jan. 2009.
- [137] D. M. Howard, G. A. Lindsey, and B. Allen, "Toward the quantification of vocal efficiency," *Journal of Voice*, vol. 4, no. 3, pp. 205–212, 1990.
- [138] A. Bouzid and N. Ellouze, "Local regularity analysis at glottal opening and closure instants in electroglottogram signal using wavelet transform modulus maxima," in *Eurospeech*, 2003, pp. 2837–2840.
- [139] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 730–743, Aug. 1986.
- [140] D. M. Howard and G. Lindsey, "Conditioned variability in voicing offsets," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 406–407, Mar. 1988.
- [141] J. C. Catford, *Fundamental Problems in Phonetics*. Indiana University Press, 1977.
- [142] Y. Lebrun and J. Hasquin-Deleval, "On the so-called 'dissociations' between electroglottogram and phonogram," *Folia Phoniatica*, vol. 23, pp. 225–227, 1971.
- [143] W. Hess and H. Indefrey, "Accurate pitch determination of speech signals by means of a laryngograph," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 9, 1984, pp. 73–76.

- [144] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, June 1992.
- [145] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.
- [146] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 7, pp. 710–732, 1992.
- [147] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 617–643, Mar. 1992.
- [148] B. M. Sadler and A. Swami, "Analysis of multiscale products for step detection and estimation," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 1043–1051, 1999.
- [149] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.
- [150] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 325–333, Sept. 1995.
- [151] J. G. McKenna, "Automatic glottal closed-phase location and analysis by Kalman filtering," in *4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Aug. 2001.
- [152] D. M. Brookes and H. P. Loke, "Modelling energy flow in the vocal tract with applications to glottal closure and opening detection," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 1999, pp. 213–216.
- [153] V. N. Tuan and C. d'Alessandro, "Robust glottal closure detection using the wavelet transform," in *Eurospeech*, Budapest, Sept. 1999, pp. 2805–2808.
- [154] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 258–265, Apr. 1994.

- [155] B. Yegnanarayana and K. S. R. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.
- [156] A. Bouzid and N. Ellouze, “Empirical mode decomposition of voiced speech signal,” in *Proc. Intl. Symp. on Control, Communications and Signal Processing*, Hammamet, Tunisia, Mar. 2004, pp. 603–606.
- [157] B. Yegnanarayana and R. Smits, “A robust method for determining instants of major excitations in voiced speech,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 1995, pp. 776–779.
- [158] P. S. Murthy and B. Yegnanarayana, “Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals,” *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 609–619, Nov. 1999.
- [159] J.-K. Chen and F. K. Soong, “An N-best candidates-based discriminative training for speech recognition applications,” *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 206–216, Jan. 1994.
- [160] M. J. Hunt, J. S. Bridle, and J. N. Holmes, “Interactive digital inverse filtering and its relation to linear prediction methods,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Tulsa, Oklahoma, Apr. 1978, pp. 15–19.
- [161] N. D. Gaubitch, “Blind identification of acoustic systems and enhancement of reverberant speech,” Ph.D. dissertation, Imperial College London, 2006.
- [162] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, “On the use of linear prediction for dereverberation of speech,” in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2003, pp. 99–102.
- [163] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

- [164] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.
- [165] J. N. Holmes, “The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer,” *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 3, pp. 298–305, 1973.
- [166] D. G. Childers and C. F. Wong, “Measuring and modeling vocal source-tract interaction,” *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 663–671, July 1994.
- [167] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
- [168] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [169] J. Hartigan and M. Wang, “A k-means clustering algorithm,” *Applied Statistics*, vol. 28, pp. 100–108, 1979.
- [170] C. Ding and X. He, “K-means clustering via principal component analysis,” in *Proc. Intl. Conf. Machine Learning (ICML)*, July 2004, pp. 225–232.
- [171] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.
- [172] Y. T. Ting, D. G. Childers, and J. C. Principe, “Tracking spectral resonances,” in *Proc. Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling*, 1988, pp. 49–54.
- [173] M. G. Jafari and M. D. Plumbley, “An adaptive orthogonal sparsifying transform for speech signals,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, Apr. 2008.
- [174] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*, 3rd ed. Springer, 1993.

- [175] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: MacMillan, 1993.
- [176] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.
- [177] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Cardiff, UK, July 2007.
- [178] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sept. 2005.
- [179] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.
- [180] M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor, "A practical multichannel dereverberation algorithm using multichannel DYPSA and spatiotemporal averaging," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2007.
- [181] M. R. P. Thomas, B. Geiser, J. Gudnason, P. A. Naylor, and P. Vary, "Voice source estimation for artificial bandwidth extension of telephone speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010.
- [182] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transforms," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.
- [183] M. R. P. Thomas, "Samples," May 2007. [Online]. Available: <http://www.commsp.ee.ic.ac.uk/~mrt102/projects/samples.html>
- [184] P. A. Naylor, "Lectures on speech processing, Imperial College London." [Online]. Available: <http://www.ee.ic.ac.uk/naylor/SpeechProcessing.html>

- [185] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, Feb. 1986, pp. 93–99.
- [186] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, no. 4, pp. 266–274, Dec. 2001.
- [187] J.-M. Valin and R. Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc. IEEE Speech Coding Workshop*, Delavan, WI, USA, Sept. 2000, pp. 130–132.
- [188] I. Uysal, H. Sathyendra, and J. G. Harris, "Bandwidth extension of telephone speech using frame-based excitation and robust features," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2005.
- [189] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 215–218.
- [190] ITU-T, *Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunications Union (ITU-T) Rec. G.711, Nov. 1998.

Appendix A

SIGMA Group Delay Window for Male and Female Speech

A.1 Varying Group Delay Window for Male Speech

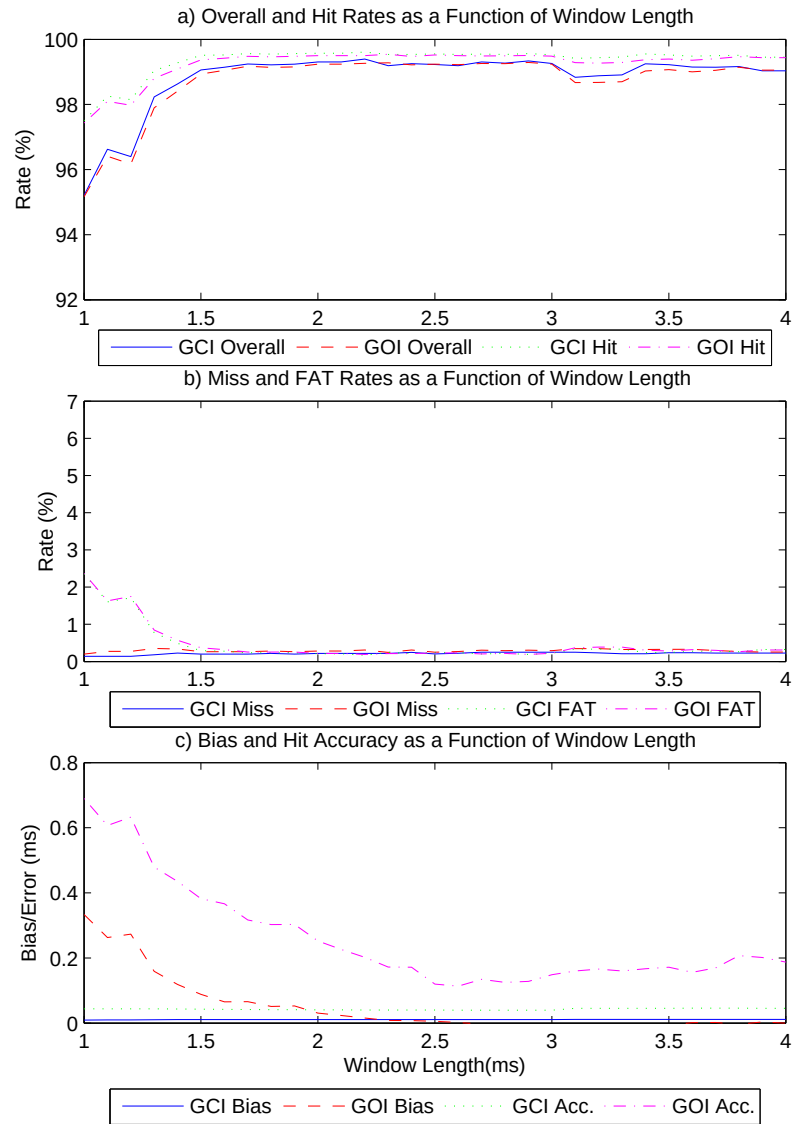


Figure A.1: SIGMA with varying group delay length for male speech.

A.2 Varying Group Delay Window for Female Speech

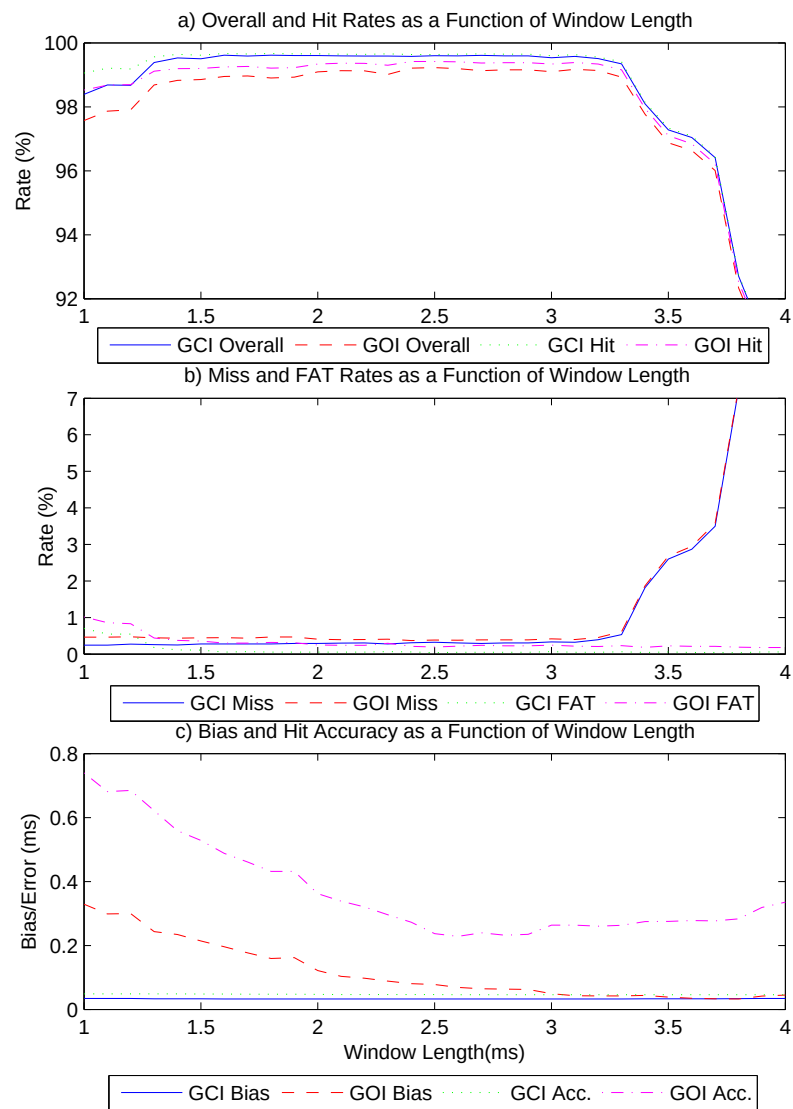


Figure A.2: SIGMA with varying group delay length for female speech.

Appendix B

GMM-Derived Voice Source Prototypes

B.1 Feature Modelling with MFCCs

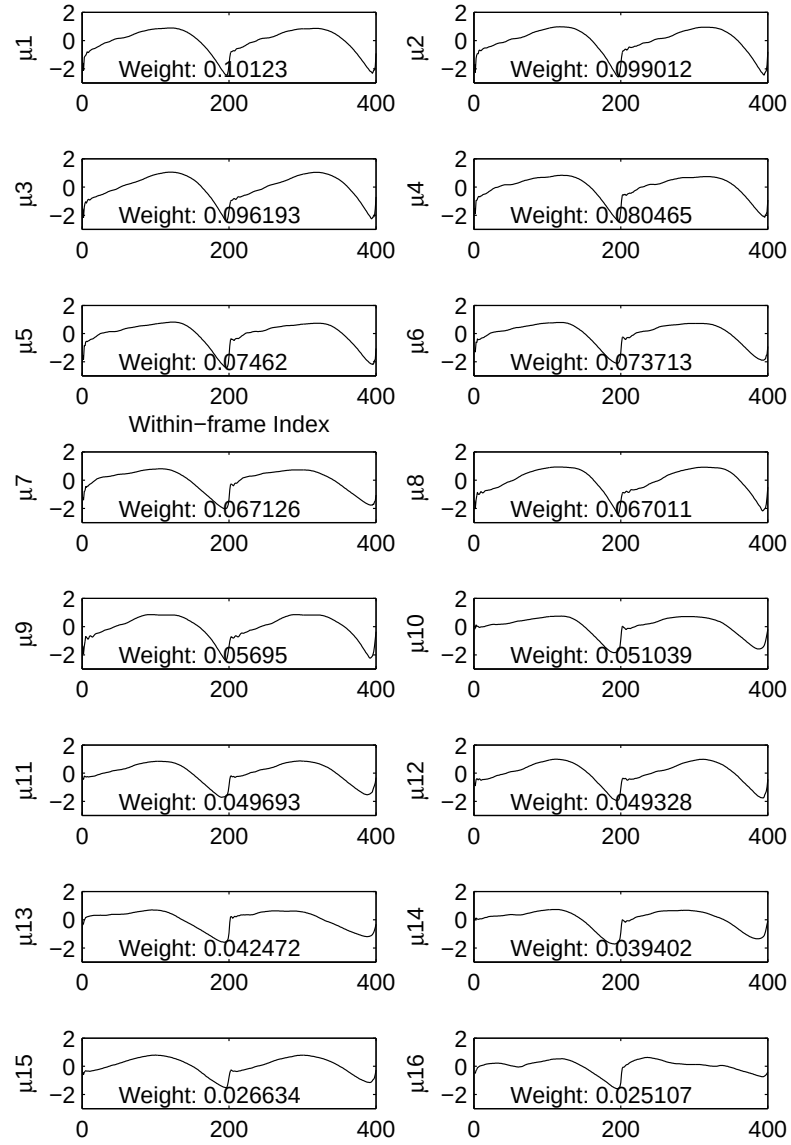


Figure B.1: Sixteen MFCC/GMM-derived voice source prototypes

B.2 Transform Modelling with PCA

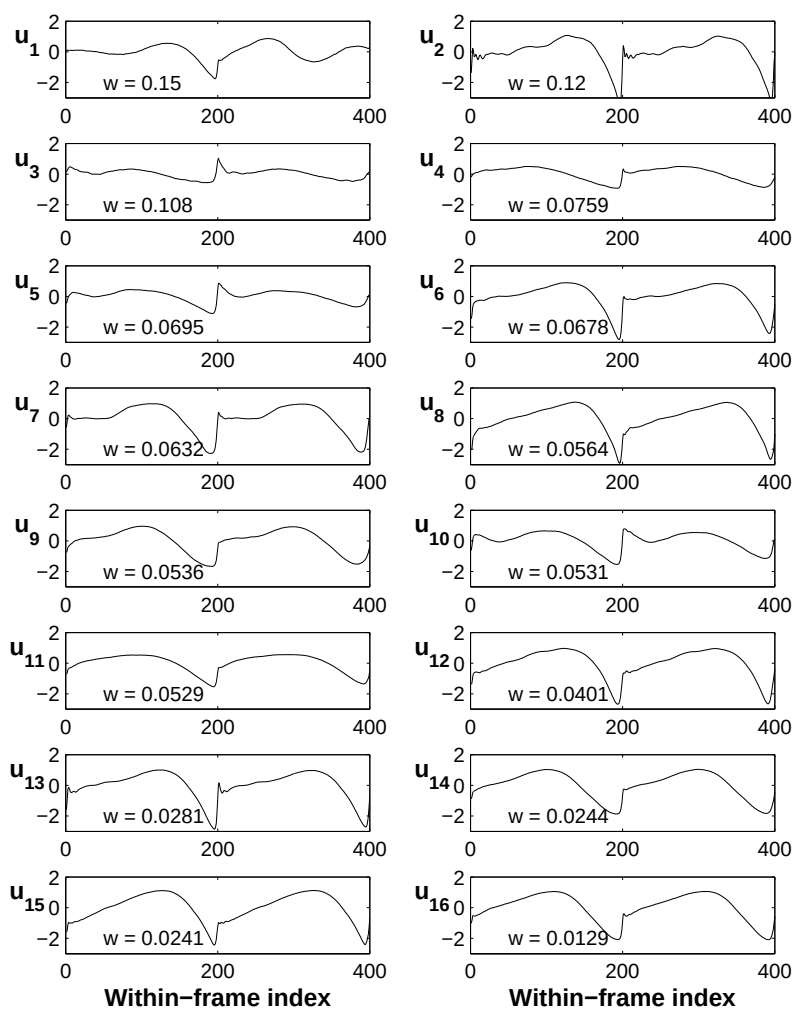


Figure B.2: Sixteen PCA/GMM-derived voice source prototypes